Building Phylogenetic Networks

Josh Collins

November 10, 2011

Recap

- At Leigh talked about a GA for clustering gene sites into differing tree topologies.

Recap

- At Leigh talked about a GA for clustering gene sites into differing tree topologies.
- Not clear how to limit complexity (number of hybridisation events).

Recap

- At Leigh talked about a GA for clustering gene sites into differing tree topologies.
- Not clear how to limit complexity (number of hybridisation events).
- Instead of a population of sets of phylogenetic trees have a population of phylogenetic networks.

Recap

- At Leigh talked about a GA for clustering gene sites into differing tree topologies.
- Not clear how to limit complexity (number of hybridisation events).
- Instead of a population of sets of phylogenetic trees have a population of phylogenetic networks.
- For crossover operator require a function capable of combining two phylogenetic networks.

Recap

- At Leigh talked about a GA for clustering gene sites into differing tree topologies.
- Not clear how to limit complexity (number of hybridisation events).
- Instead of a population of sets of phylogenetic trees have a population of phylogenetic networks.
- For crossover operator require a function capable of combining two phylogenetic networks.
- For GAs don't want to be solving $\mathcal{NP}$-hard problems as part of algorithm.

Recap

- At Leigh talked about a GA for clustering gene sites into differing tree topologies.
- Not clear how to limit complexity (number of hybridisation events).
- Instead of a population of sets of phylogenetic trees have a population of phylogenetic networks.
- For crossover operator require a function capable of combining two phylogenetic networks.
- For GAs don't want to be solving $\mathcal{NP}$-hard problems as part of algorithm.
- Want to ideally be at least $\mathcal{P}$.

Recap

Aim

Construct a phylogenetic network that displays a pair of phylogenetic networks.

Recap

Aim

Construct a phylogenetic network that displays a pair of phylogenetic networks.

Aim'

Construct a phylogenetic network that displays a set of phylogenetic trees.

Maximal Acyclic Agreement Forests

- An acyclic agreement forest $\mathcal{F}$ for a set of phylogenetic trees $\mathscr{T}$ is a collection of trees such that
  - Each tree in $\mathcal{F}$ is a subtree of all trees in $\mathscr{T}$,
  - No pair of trees in $\mathcal{F}$ overlap in any tree in $\mathscr{T}$.
  - For a pair of trees $\mathcal{T}_i, \mathcal{T}_j \in \mathcal{F}$ if there is a path in a tree in $\mathscr{T}$ from the most recent common ancestor (MRCA) of $\mathcal{T}_i$ to the MRCA of $\mathcal{T}_j$ then there is no tree in $\mathscr{T}$ such that there is a path in the opposite direction.

Maximal Acyclic Agreement Forests

- An acyclic agreement forest $\mathcal{F}$ for a set of phylogenetic trees $\mathscr{T}$ is a collection of trees such that
  - Each tree in $\mathcal{F}$ is a subtree of all trees in $\mathscr{T}$,
  - No pair of trees in $\mathcal{F}$ overlap in any tree in $\mathscr{T}$.
  - For a pair of trees $\mathcal{T}_i, \mathcal{T}_j \in \mathcal{F}$ if there is a path in a tree in $\mathscr{T}$ from the most recent common ancestor (MRCA) of $\mathcal{T}_i$ to the MRCA of $\mathcal{T}_j$ then there is no tree in $\mathscr{T}$ such that there is a path in the opposite direction.
- Going to make use of forests as a way to create an upper bound on hybridisation events of constructed network.

Maximal Acyclic Agreement Forests

- An acyclic agreement forest $\mathcal{F}$ for a set of phylogenetic trees $\mathscr{T}$ is a collection of trees such that
  - Each tree in $\mathcal{F}$ is a subtree of all trees in $\mathscr{T}$,
  - No pair of trees in $\mathcal{F}$ overlap in any tree in $\mathscr{T}$.
  - For a pair of trees $\mathcal{T}_i, \mathcal{T}_j \in \mathcal{F}$ if there is a path in a tree in $\mathscr{T}$ from the most recent common ancestor (MRCA) of $\mathcal{T}_i$ to the MRCA of $\mathcal{T}_j$ then there is no tree in $\mathscr{T}$ such that there is a path in the opposite direction.
- Going to make use of forests as a way to create an upper bound on hybridisation events of constructed network.
- Calculating a maximum acyclic agreement forest for a pair of trees is $\mathcal{NP}$-hard. (Bordewich & Semple, 2005)

Maximal Acyclic Agreement Forests

- An acyclic agreement forest $\mathcal{F}$ for a set of phylogenetic trees $\mathscr{T}$ is a collection of trees such that
  - Each tree in $\mathcal{F}$ is a subtree of all trees in $\mathscr{T}$,
  - No pair of trees in $\mathcal{F}$ overlap in any tree in $\mathscr{T}$.
  - For a pair of trees $\mathcal{T}_i, \mathcal{T}_j \in \mathcal{F}$ if there is a path in a tree in $\mathscr{T}$ from the most recent common ancestor (MRCA) of $\mathcal{T}_i$ to the MRCA of $\mathcal{T}_j$ then there is no tree in $\mathscr{T}$ such that there is a path in the opposite direction.
- Going to make use of forests as a way to create an upper bound on hybridisation events of constructed network.
- Calculating a maximum acyclic agreement forest for a pair of trees is $\mathcal{NP}$-hard. (Bordewich & Semple, 2005)
- Is there a weakening of maximum acyclic agreement forests such that one can be calculated in $\mathcal{P}$ time?

Maximal Acyclic Agreement Forests

Definition

Let $\mathcal{F}$ and $\mathcal{F}'$ be two forests. Call $\mathcal{F}'$ a *subforest* of $\mathcal{F}$ if the members of $\mathcal{F}'$ are non-overlapping subtrees of $\mathcal{F}$.

Maximal Acyclic Agreement Forests

Definition

Let $\mathcal{F}$ and $\mathcal{F}'$ be two forests. Call $\mathcal{F}'$ a *subforest* of $\mathcal{F}$ if the members of $\mathcal{F}'$ are non-overlapping subtrees of $\mathcal{F}$.

Theorem

The above definition puts a partial ordering on the set of phylogenetic forests.

Maximal Acyclic Agreement Forests

Definition

Let $\mathcal{F}$ and $\mathcal{F}'$ be two forests. Call $\mathcal{F}'$ a *subforest* of $\mathcal{F}$ if the members of $\mathcal{F}'$ are non-overlapping subtrees of $\mathcal{F}$.

Theorem

The above definition puts a partial ordering on the set of phylogenetic forests.

Definition

A forest is *maximal* (with respect to a set of phylogenetic trees $\mathscr{T}$) if it is an acyclic forest for $\mathscr{T}$ and any subforest of $\mathcal{F}$ that is also an acyclic forest for $\mathscr{T}$ is $\mathcal{F}$.

Maximal Acyclic Agreement Forests

Definition

Let $\mathcal{F}$ and $\mathcal{F}'$ be two forests. Call $\mathcal{F}'$ a *subforest* of $\mathcal{F}$ if the members of $\mathcal{F}'$ are non-overlapping subtrees of $\mathcal{F}$.

Theorem

The above definition puts a partial ordering on the set of phylogenetic forests.

Definition

A forest is *maximal* (with respect to a set of phylogenetic trees $\mathscr{T}$) if it is an acyclic forest for $\mathscr{T}$ and any subforest of $\mathcal{F}$ that is also an acyclic forest for $\mathscr{T}$ is $\mathcal{F}$.

Theorem

Calculating a maximal acyclic agreement forest takes polynomial time.

$\mathcal{F} \leftarrow \{\text{'}\ell\text{'} : \ell \in L(\mathscr{T})\}$; $\{\ell_1, \ell_2, \ldots, \ell_n\} \leftarrow L(\mathcal{F})$;
for $i = 1, \ldots, n$ **do**

  for $j = i + 1, \ldots, n$ **do**
    $\mathcal{T}_1 \leftarrow$ The tree $\mathcal{T}_1 \in \mathcal{F}$ such that $\ell_i \in L(\mathcal{T}_1)$;
    $\mathcal{T}_2 \leftarrow$ The tree $\mathcal{T}_2 \in \mathcal{F}$ such that $\ell_j \in L(\mathcal{T}_2)$;
    $L_1 \leftarrow L(\mathcal{T}_1)$; $L_2 \leftarrow L(\mathcal{T}_2)$;
    $\mathcal{T} :\in \mathscr{T}$; $\mathcal{T}' \leftarrow \mathcal{T} \mid (L_1 \cup L_2)$;
    if $\neg$ isSubtree$(\mathcal{T}_1, \mathcal{T}_2, \mathscr{T})$ **then continue**;
    if isOverlap$(\mathcal{T}', \mathcal{F} \setminus \{\mathcal{T}_1, \mathcal{T}_2\}, \mathscr{T})$ **then continue**;
    if IsCyclic$(\mathcal{T}', \mathcal{F} \setminus \{\mathcal{T}_1, \mathcal{T}_2\}, \mathscr{T})$ **then continue**;
    else $\mathcal{F} \leftarrow \{\mathcal{T}'\} \cup \mathcal{F} \setminus \{\mathcal{T}_1, \mathcal{T}_2\}$;
  end
end
return $\mathcal{F}$;

Function findMaximalForest$(\mathscr{T})$

How long does it take to find a maximal acyclic agreement forest?

$$\textsc{findMaximalForest}(\mathscr{T}) \in \mathcal{O}(|X|^5 \, |\mathscr{T}|)$$

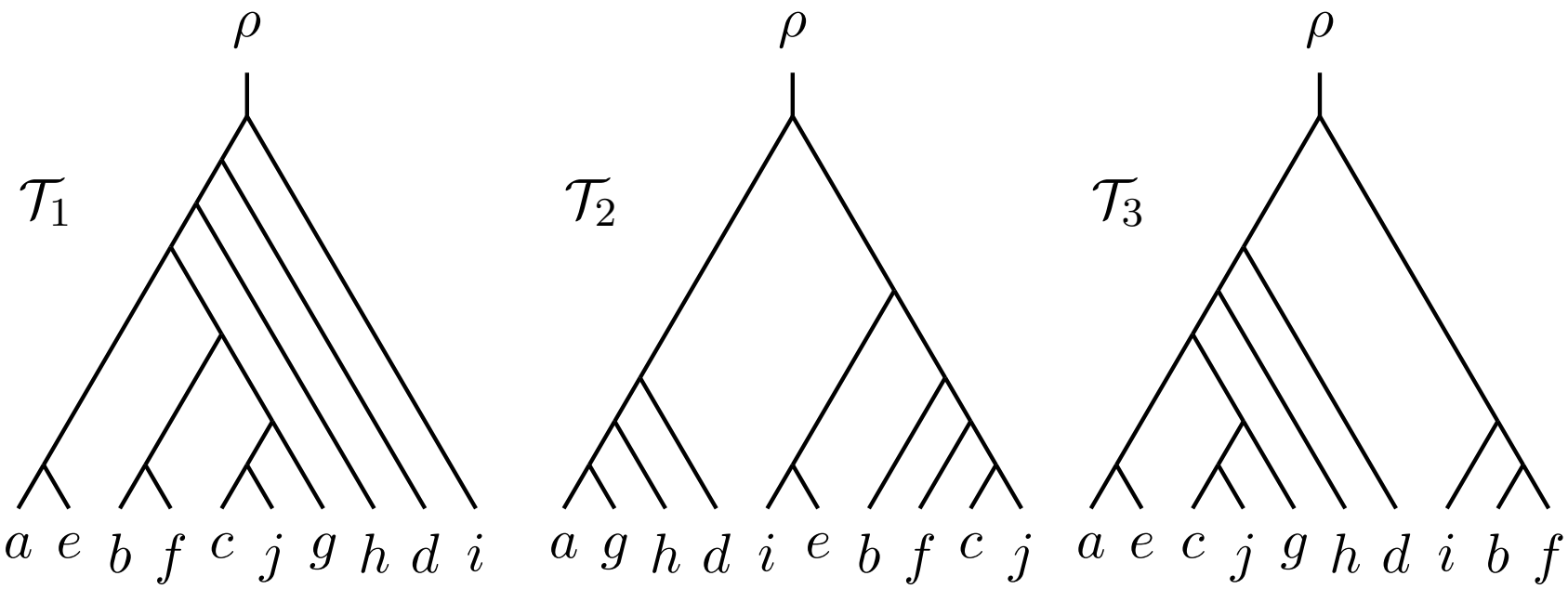
Construction

- Input now consists of a set of phylogenetic trees $\mathscr{T}$ and a phylogenetic forest $\mathcal{F}$.

Construction

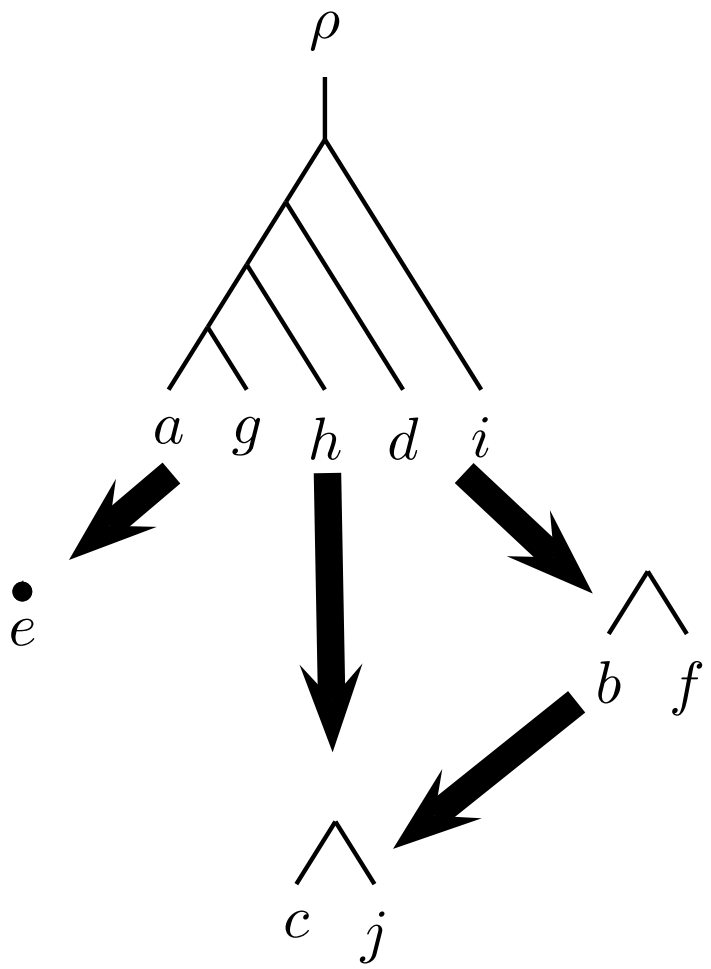
- Input now consists of a set of phylogenetic trees $\mathscr{T}$ and a phylogenetic forest $\mathcal{F}$.
- Output a phylogenetic network $\mathcal{N}$ such that
  - $h(\mathscr{T}) \leq h(\mathcal{N}) \leq |\mathcal{F}|\,|\mathscr{T}|$
  - $\mathcal{N}$ displays $\mathscr{T}$, that is there is some deletion of the egdes of $\mathcal{N}$ that gives $\mathcal{T}$ for each $\mathcal{T} \in \mathscr{T}$.
  - The forest induced by $\mathcal{N}$ is a superforest of $\mathcal{F}$, that is $\mathcal{F}$ is a subforest of the forest obtained by deleting all the in-arcs of vertices with in-degree greater than or equal to two.

Example

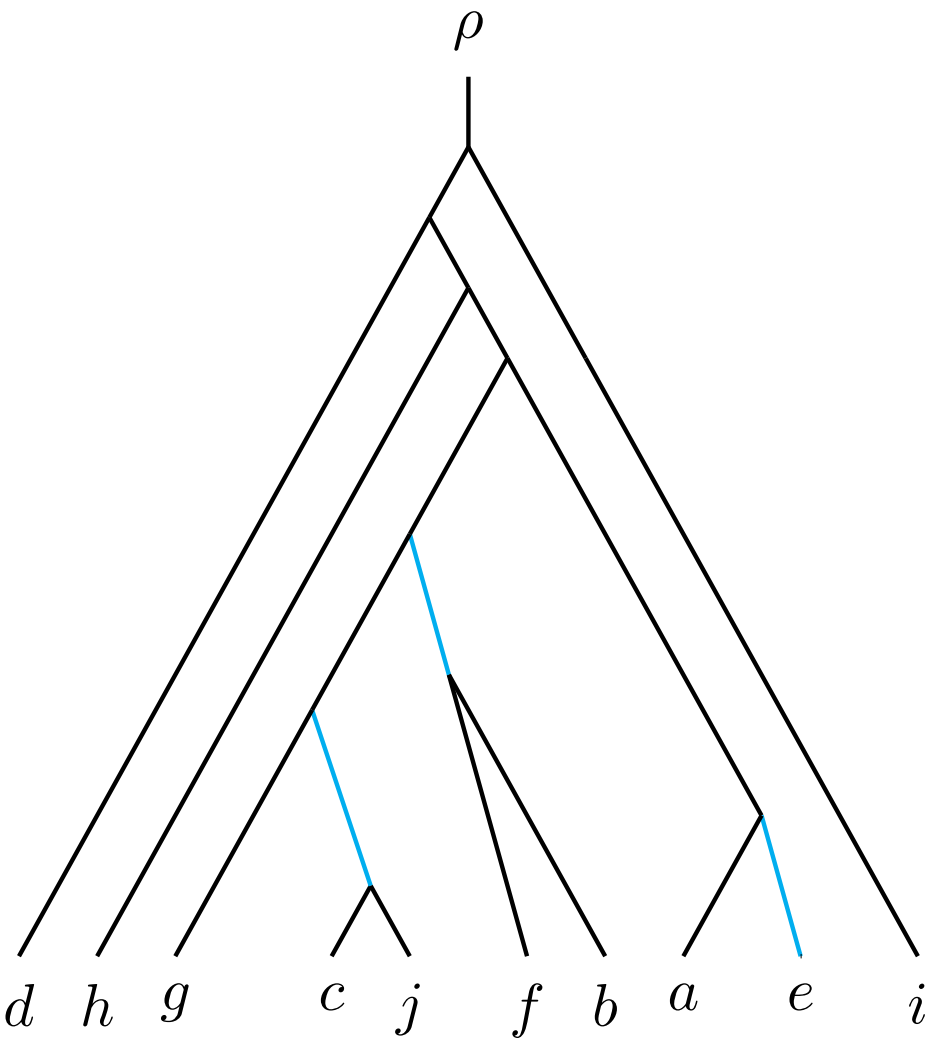


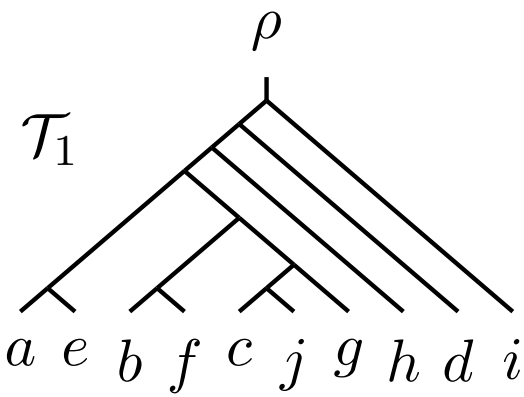
$$\mathcal{F} = \left\{ \rho \atop a\ g\ h\ d\ i \;,\; \wedge \atop b\ f \;,\; \wedge \atop c\ j \;,\; \bullet \atop e \right\}$$

Example

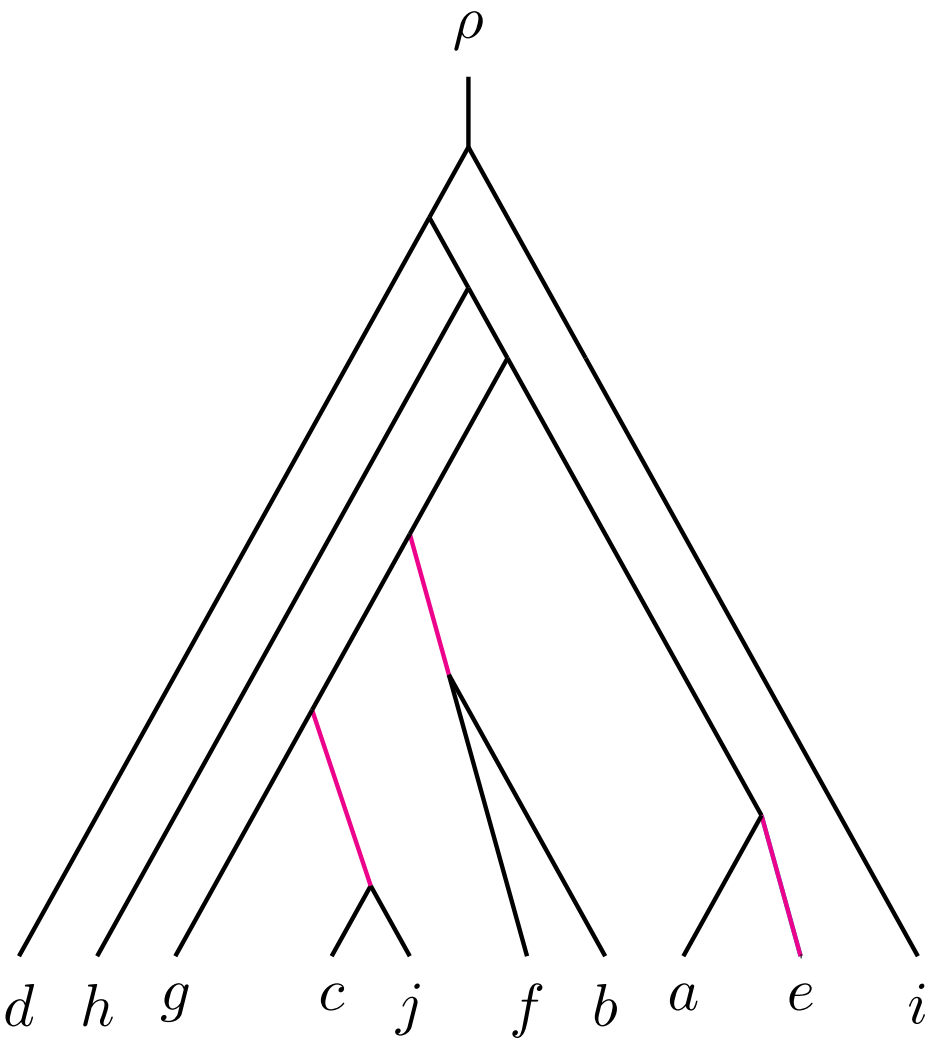


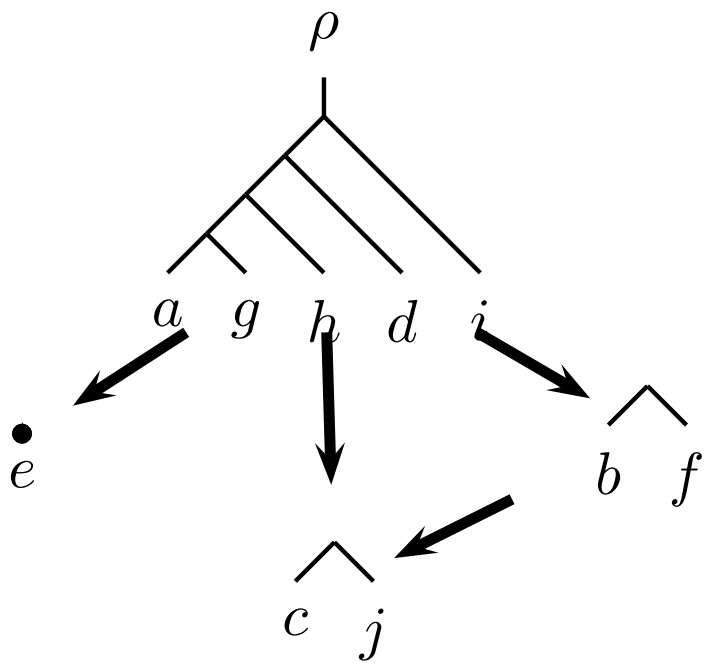
Example

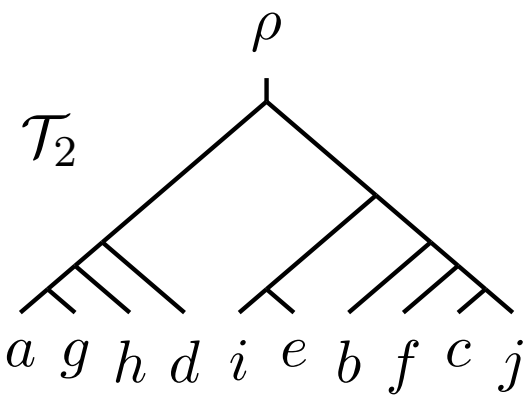




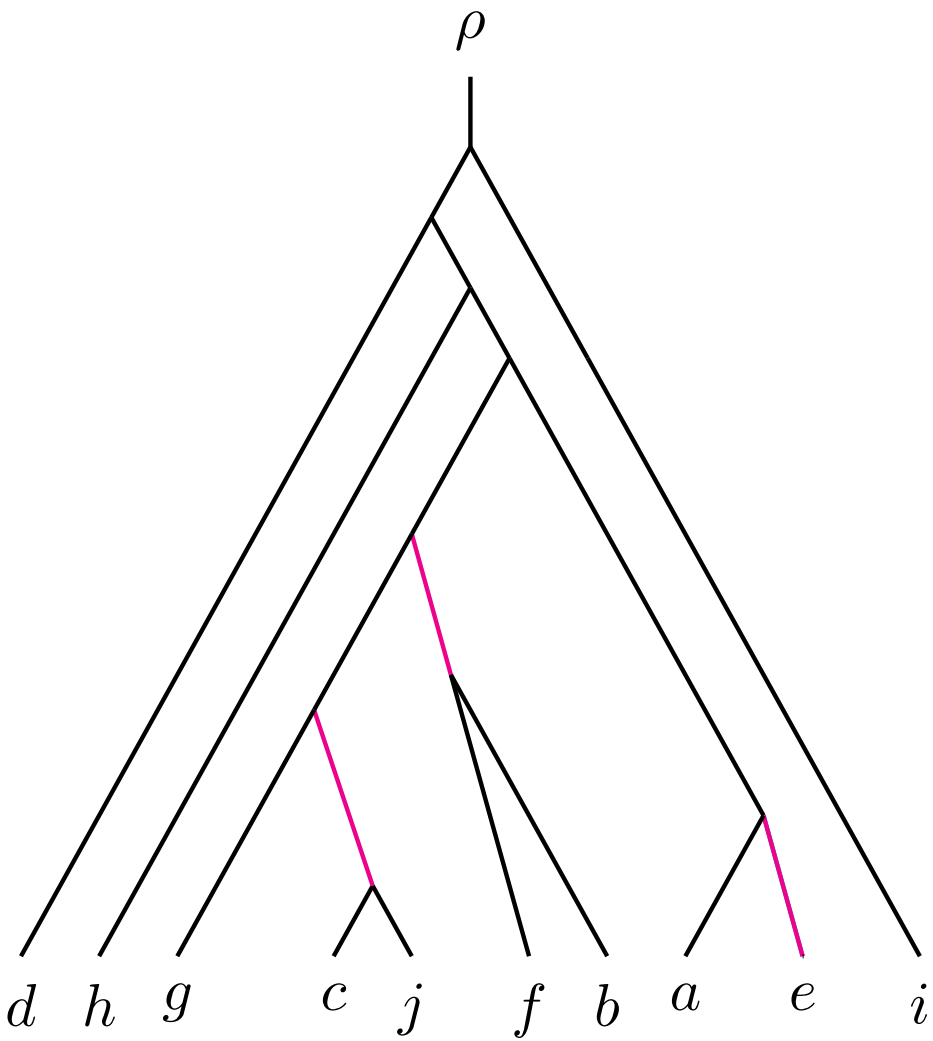
Example

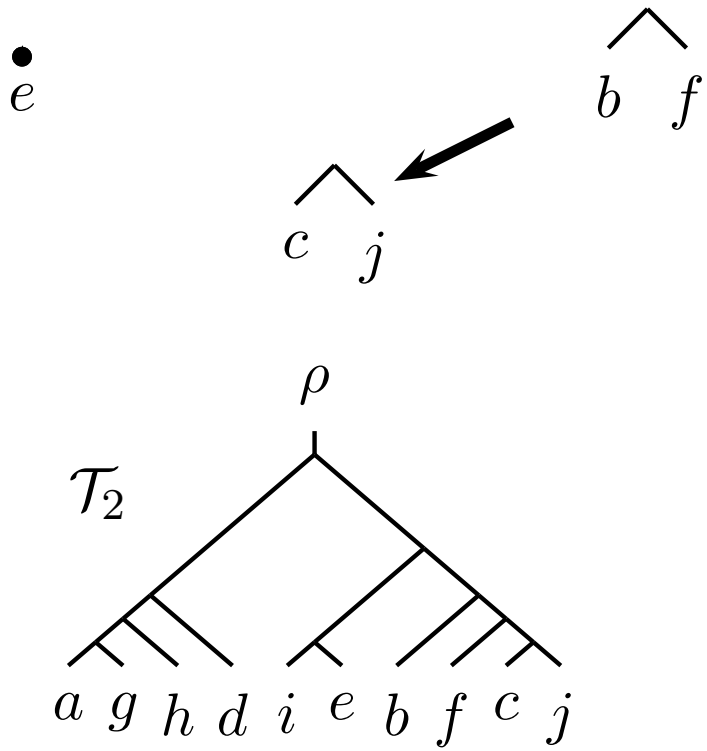




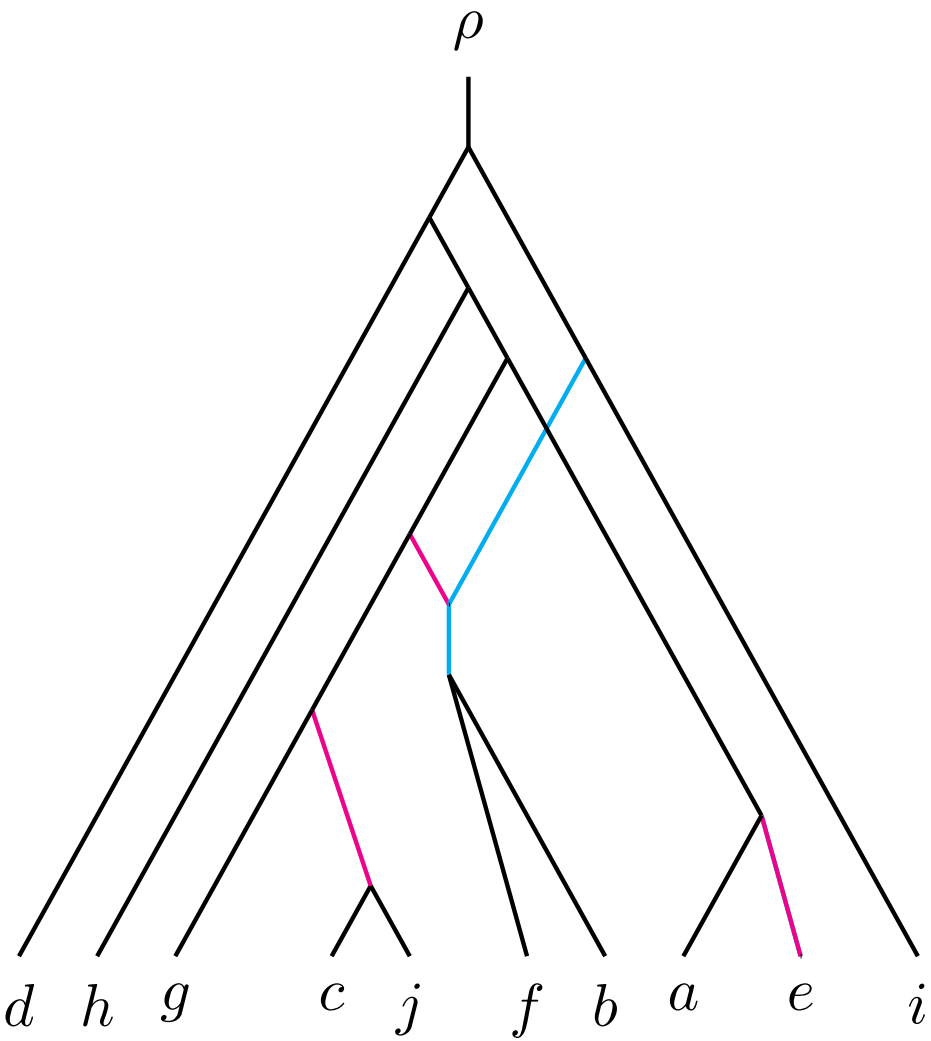


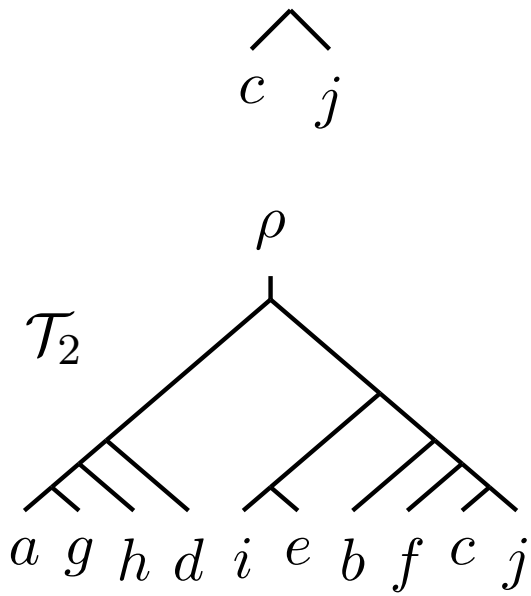
Example



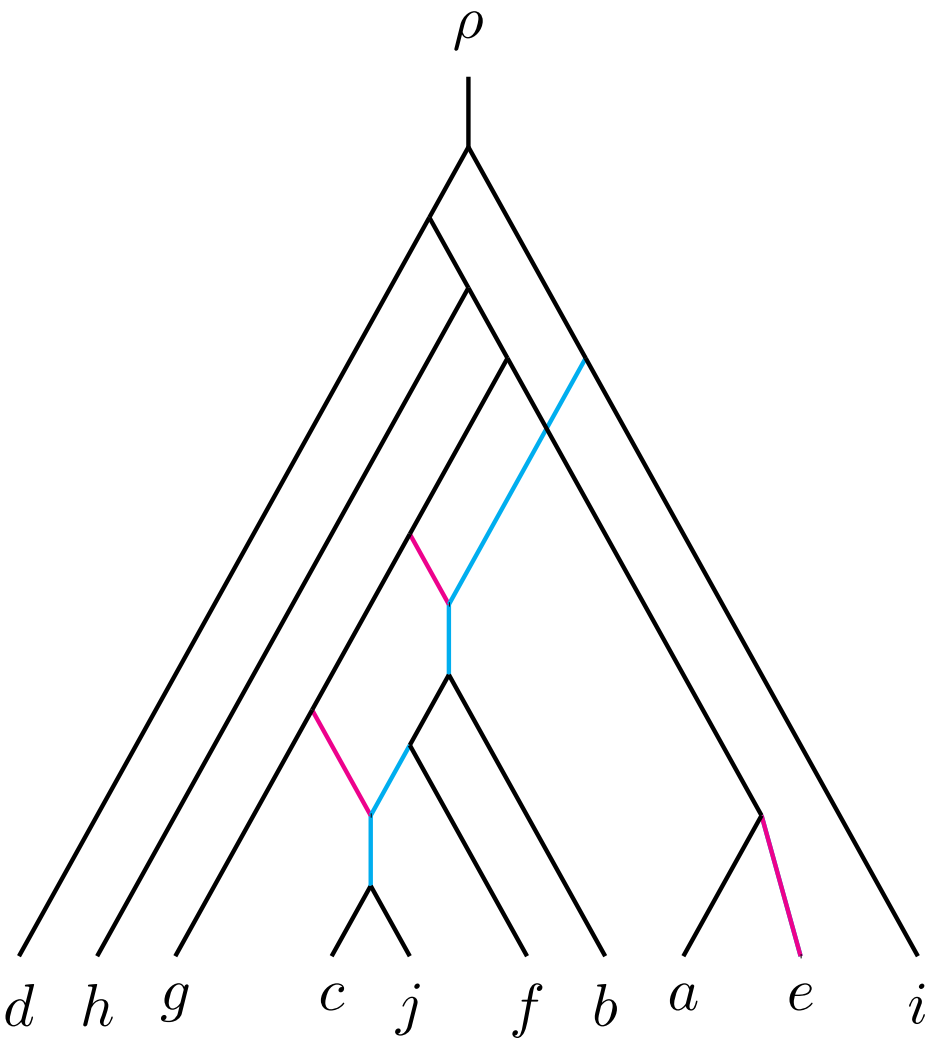


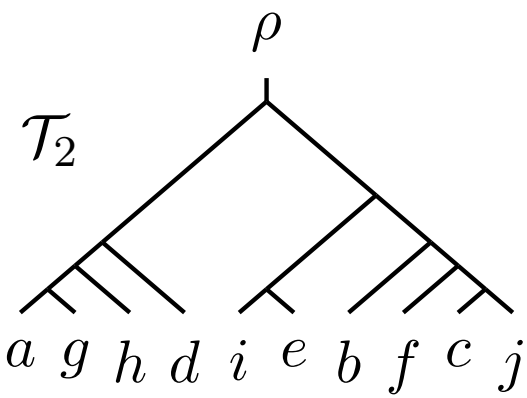
Example



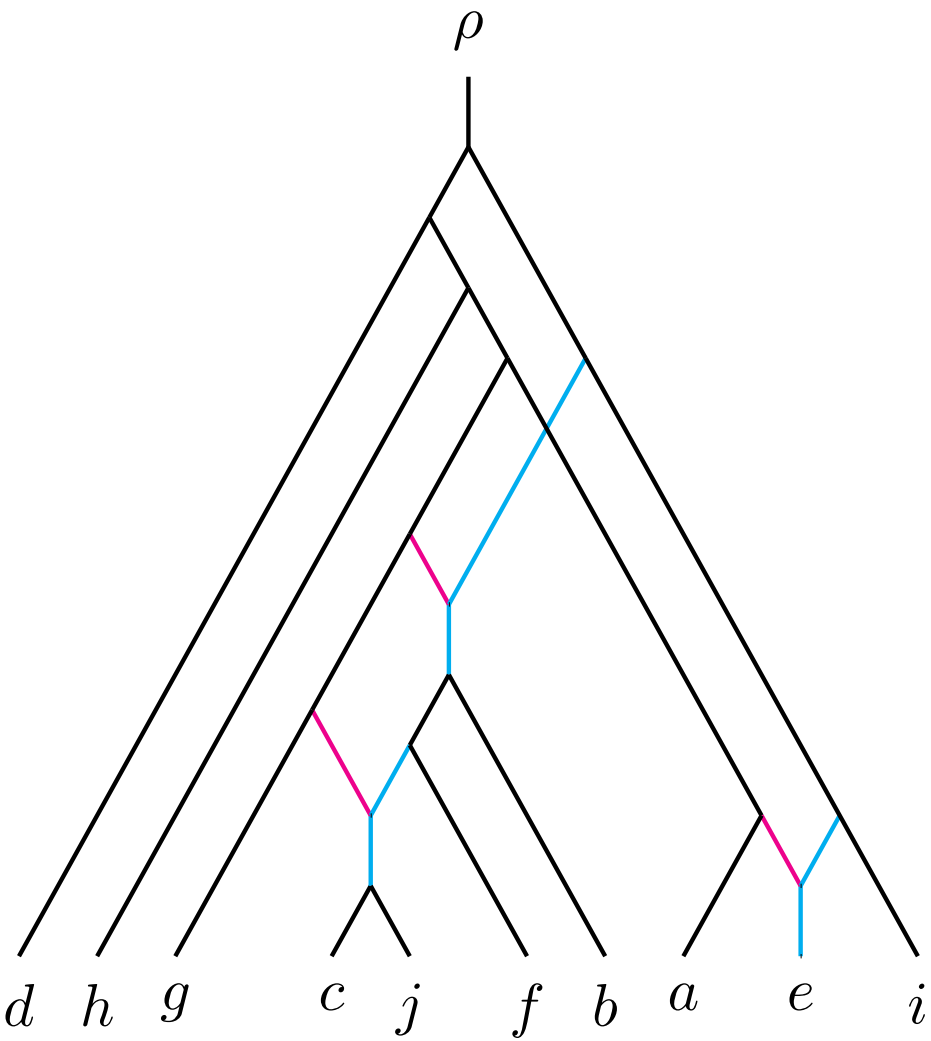


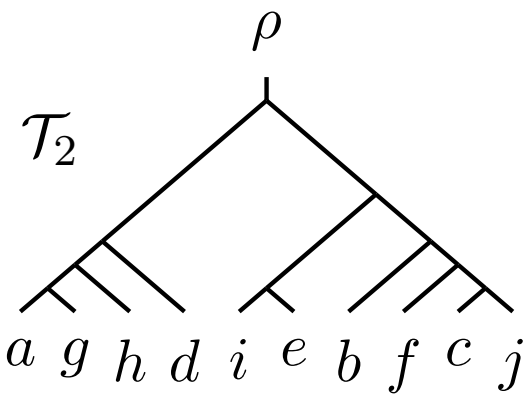
Example



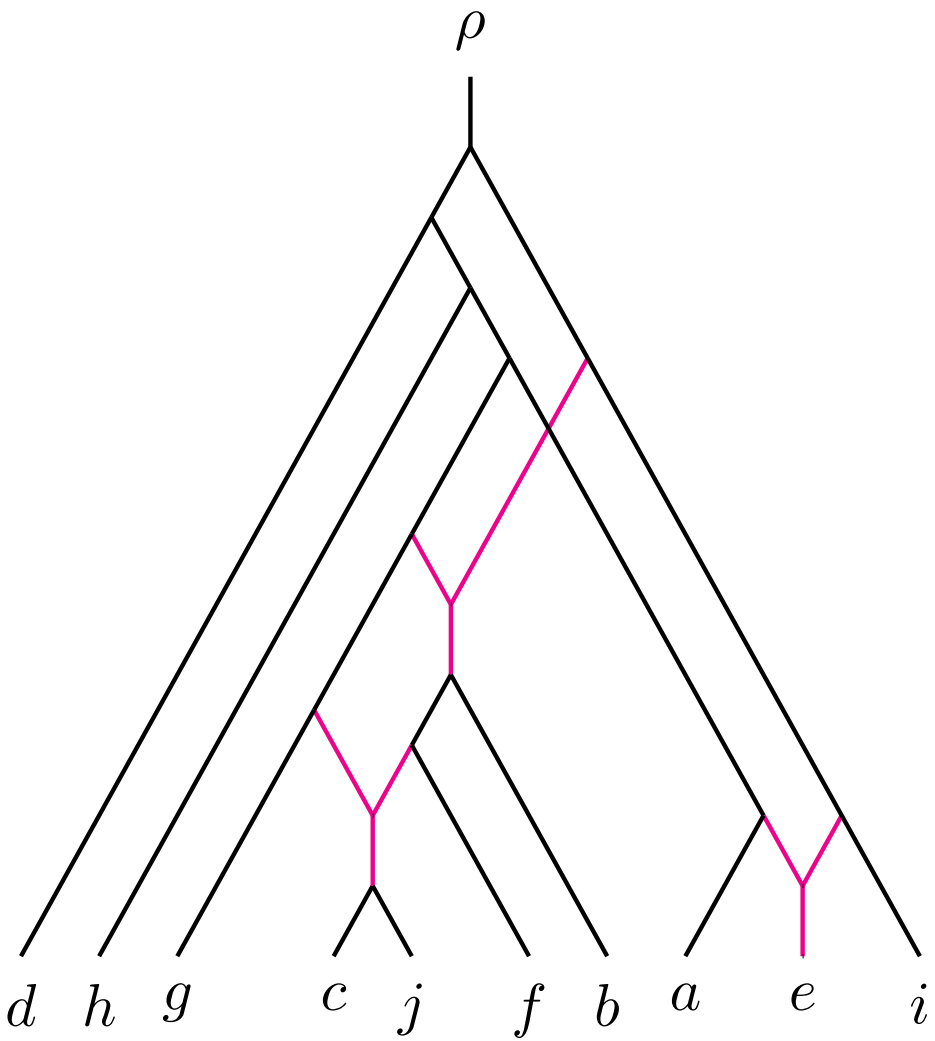


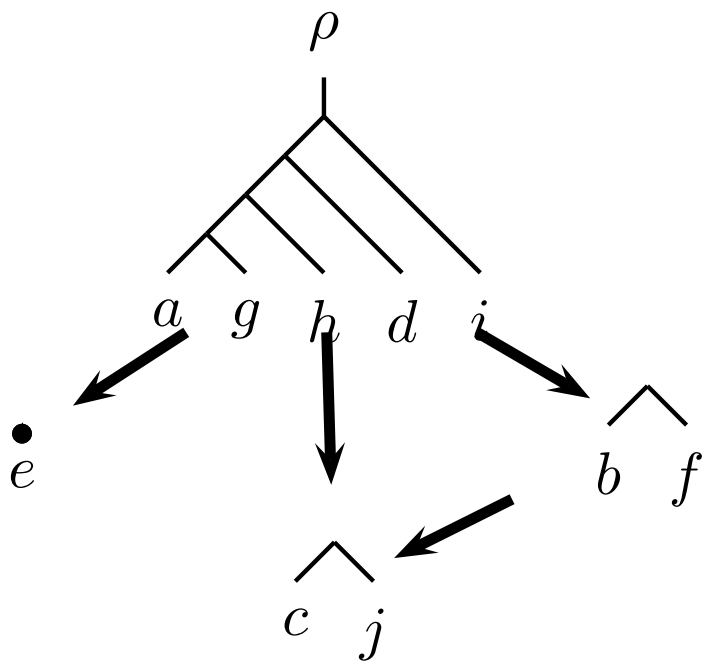
Example

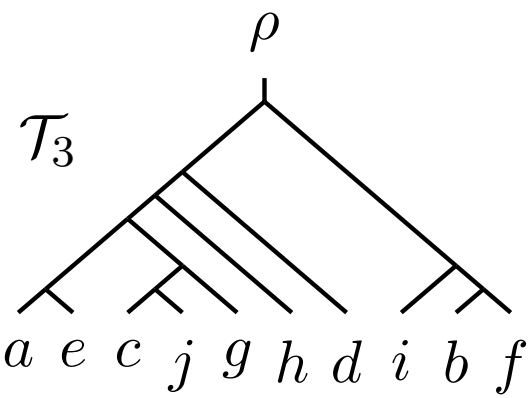




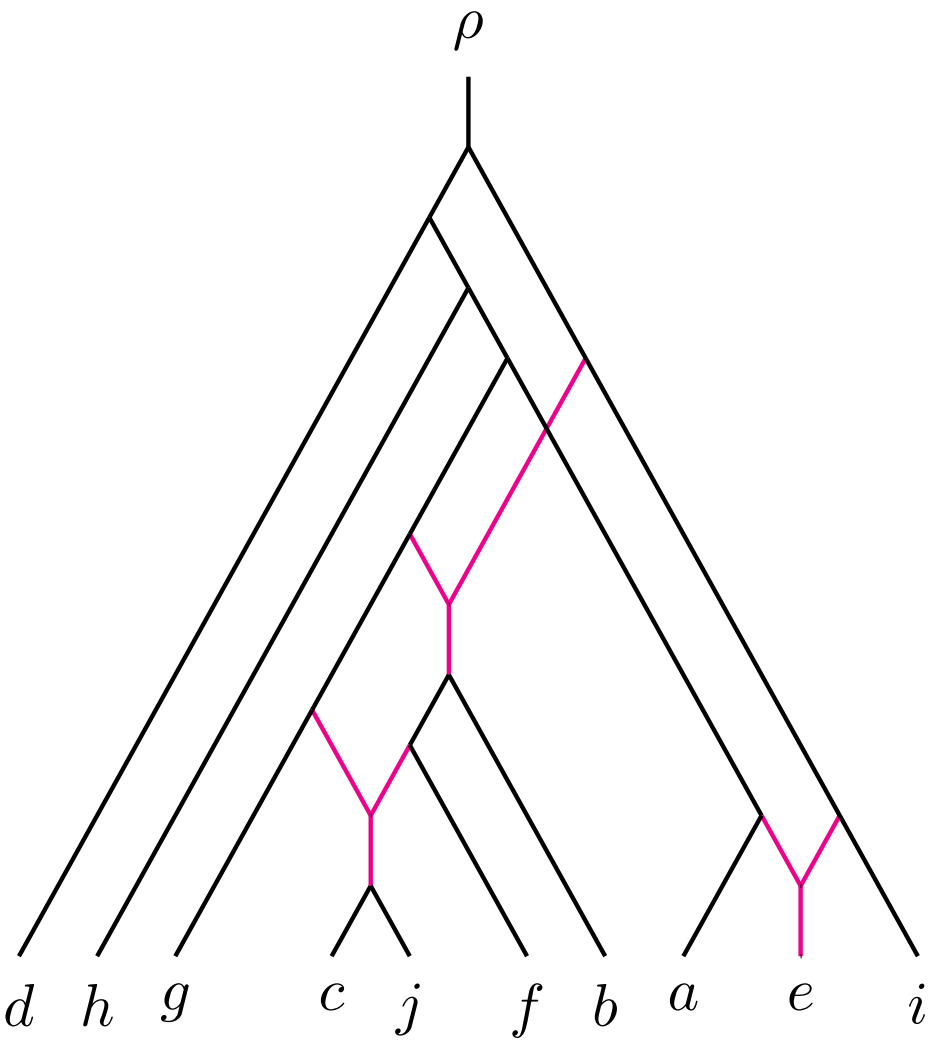
Example

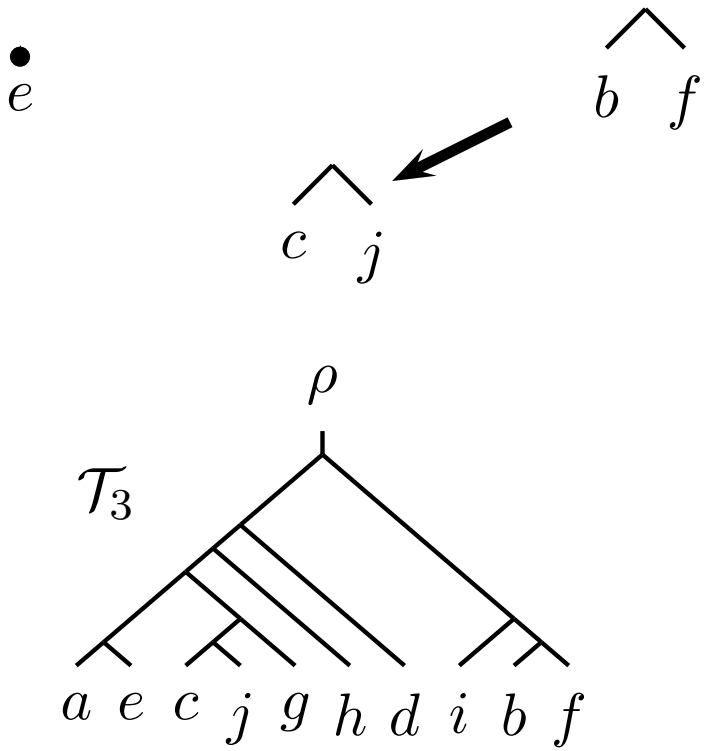




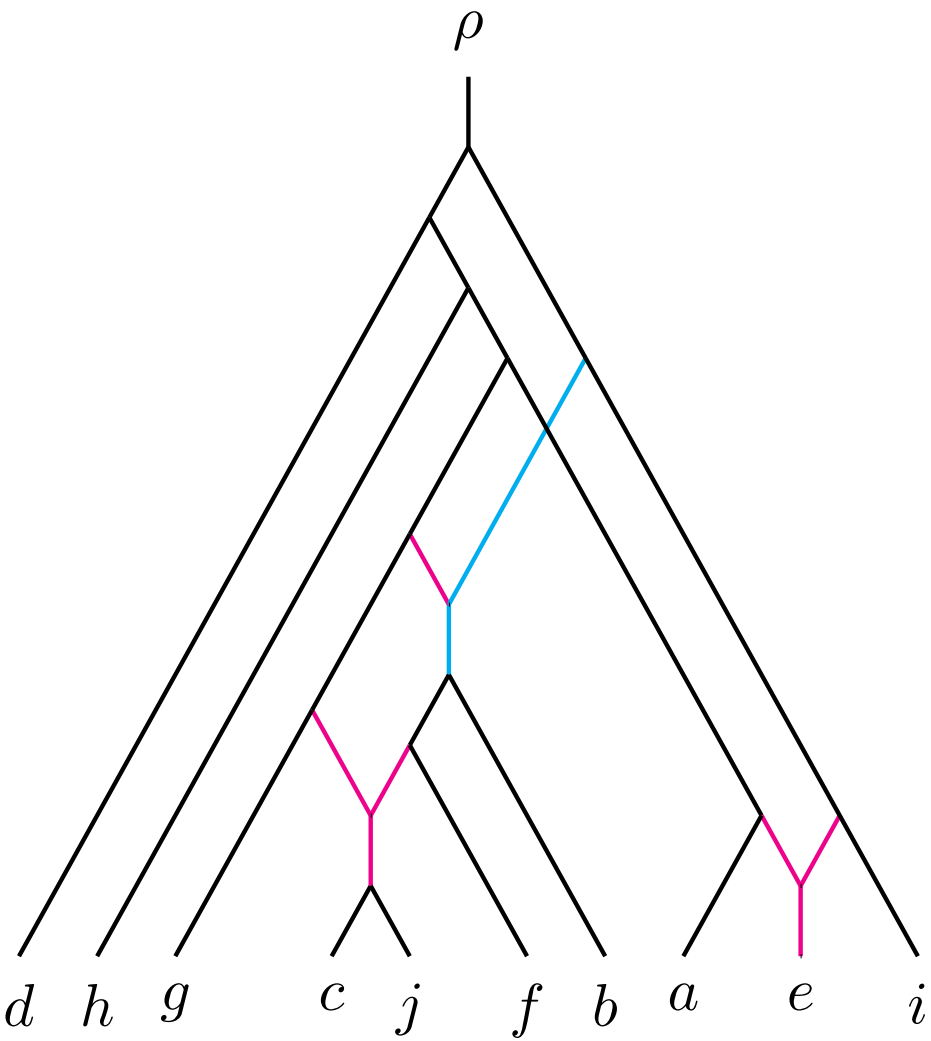


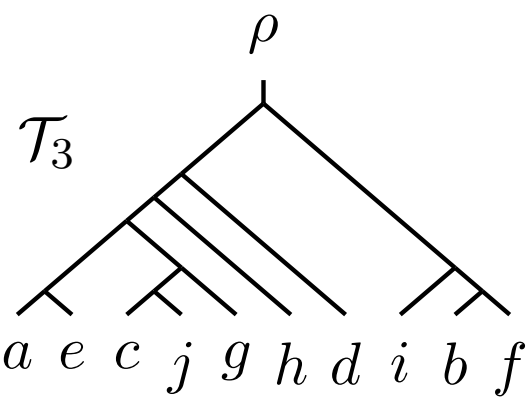
Example



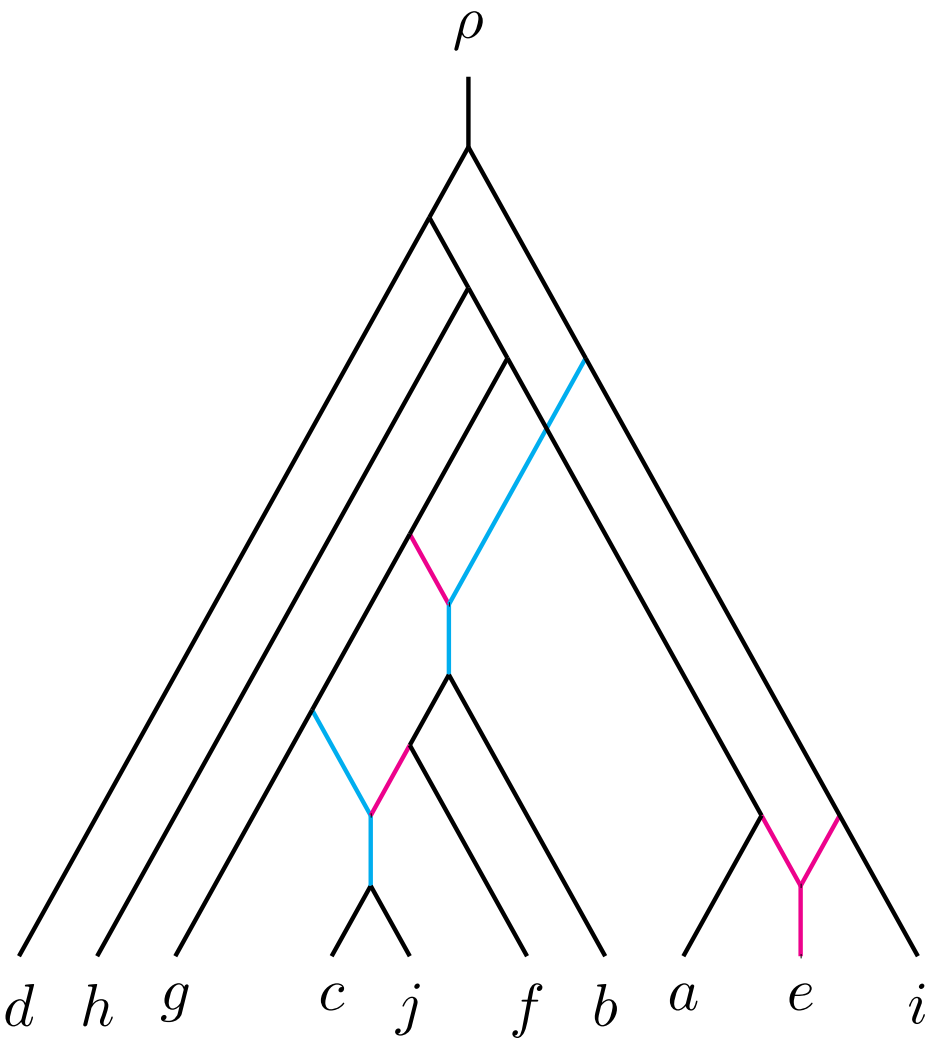


Example

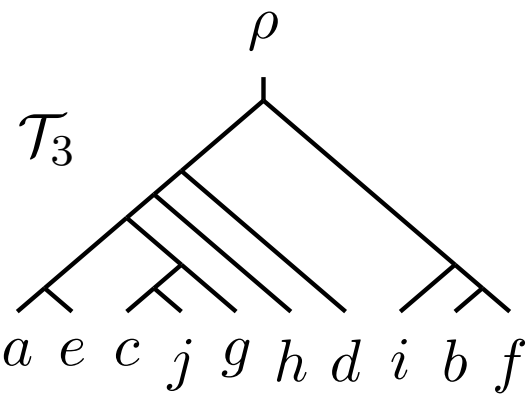




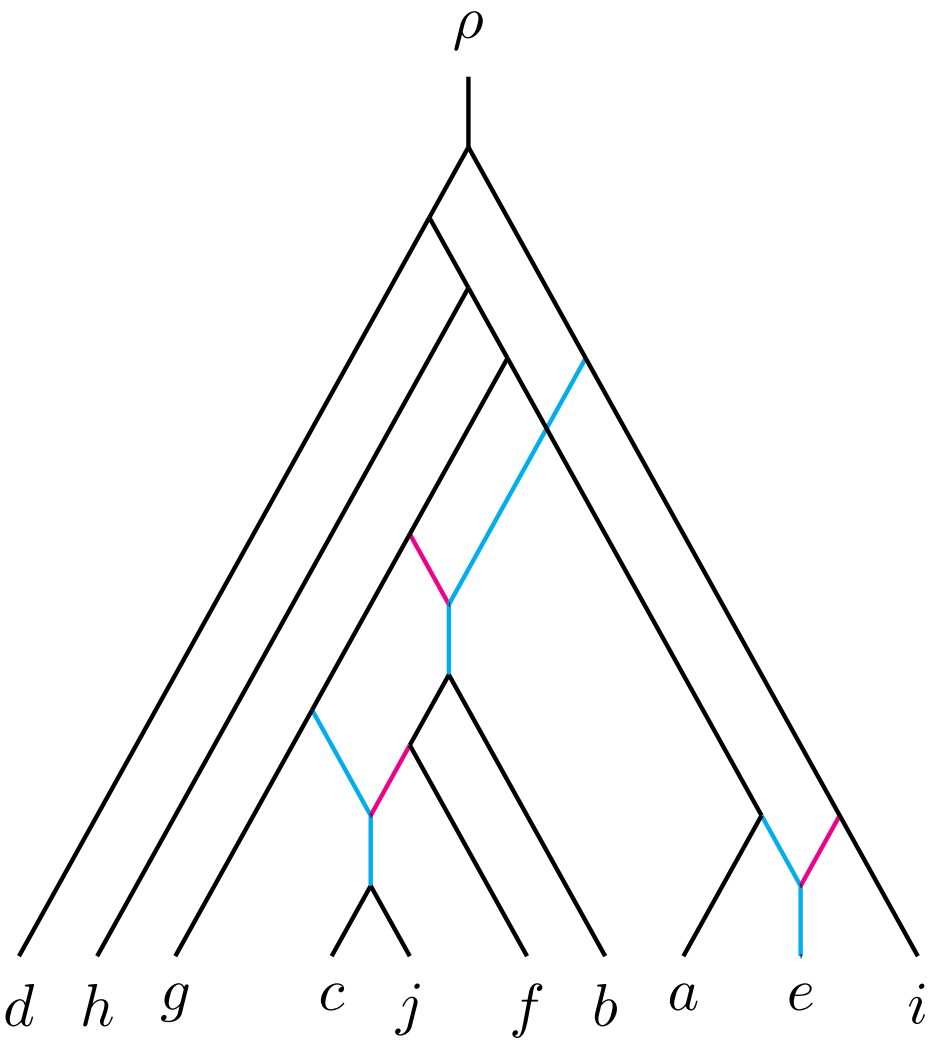
Example

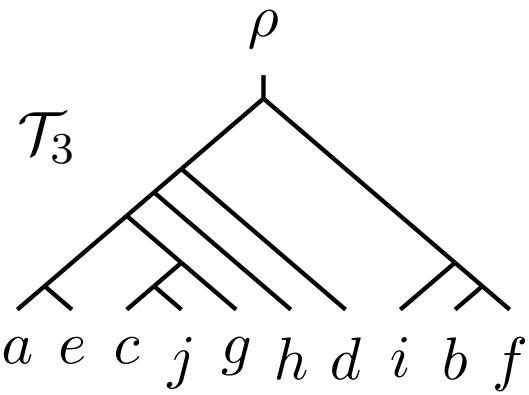


e

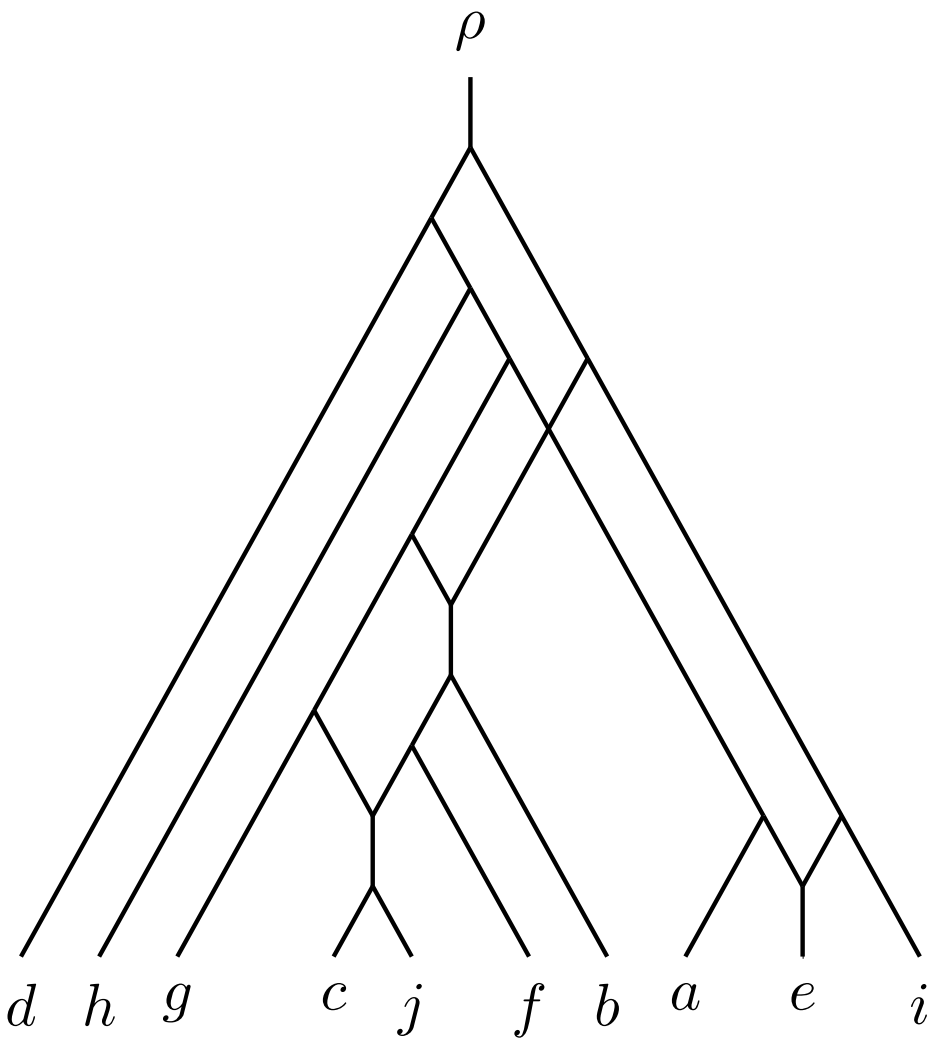


Example





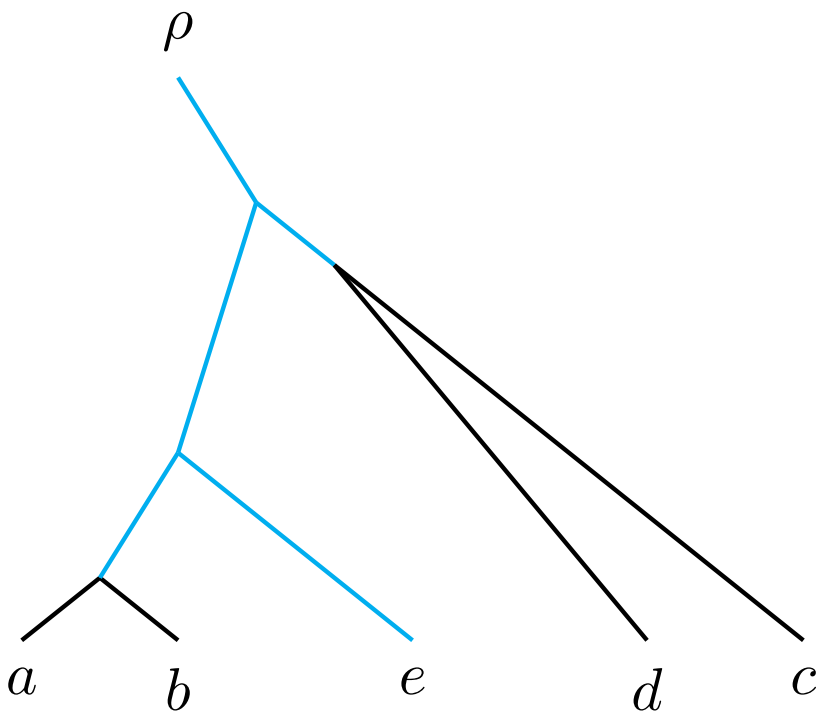
Example

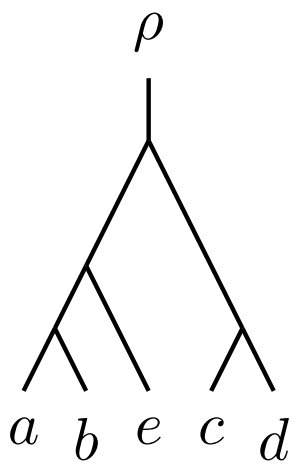


Great, so what's the problem?

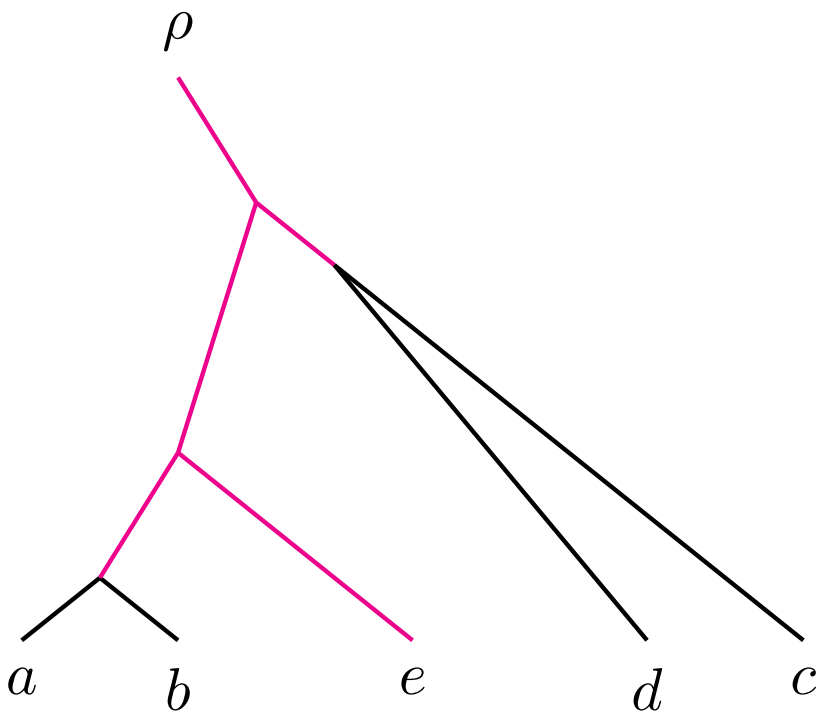
$$\mathscr{T} = \left\{ \begin{array}{ccccc} \rho & & \rho & & \rho \\ \text{tree 1} & , & \text{tree 2} & , & \text{tree 3} \\ a\ b\ e\ c\ d & & a\ b\ c\ e\ d & & a\ c\ d\ b\ e \end{array} \right\}$$

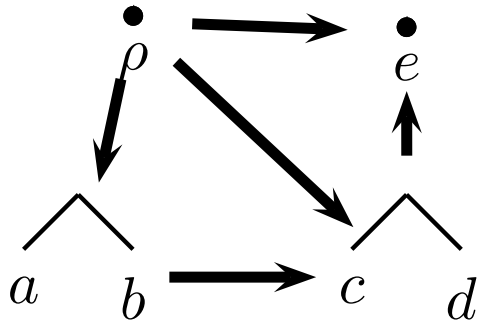
Great, so what's the problem?

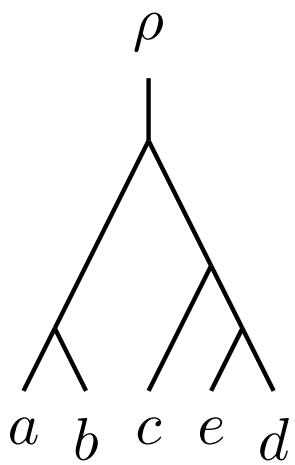




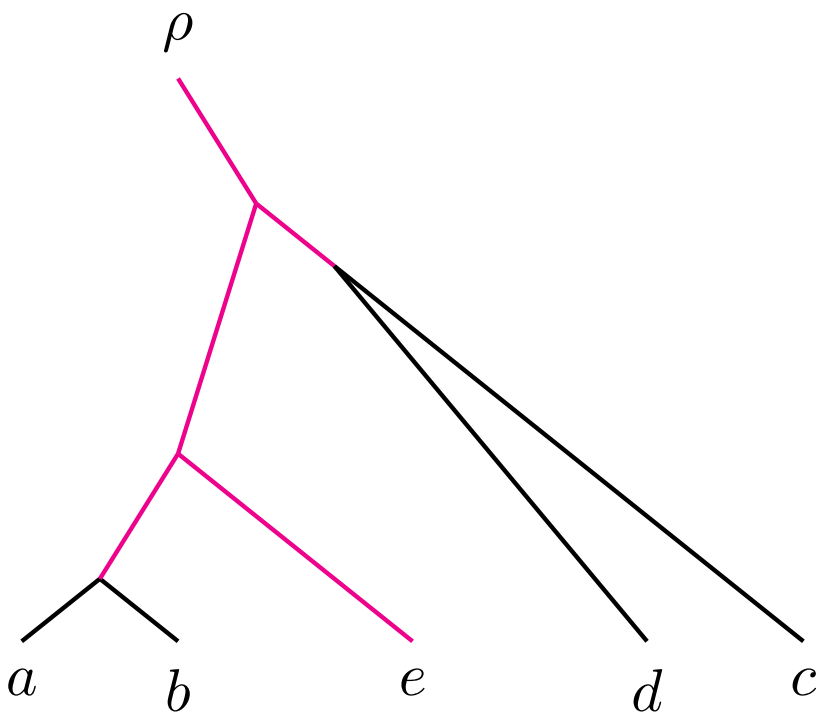
Great, so what's the problem?

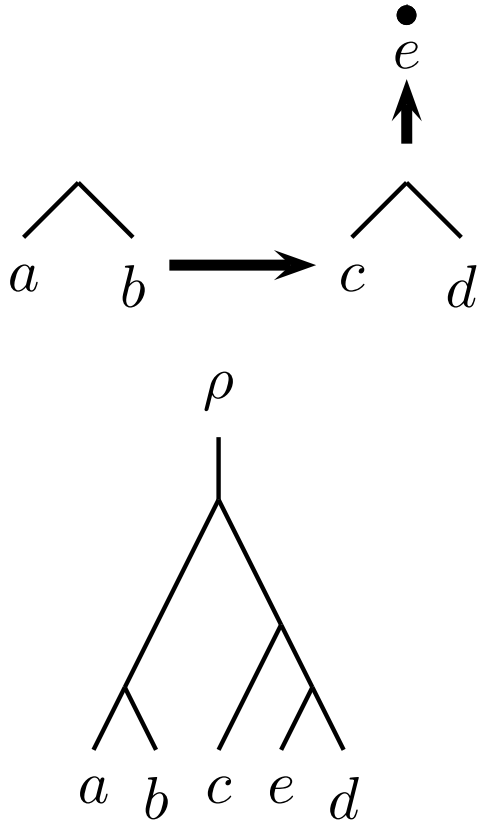




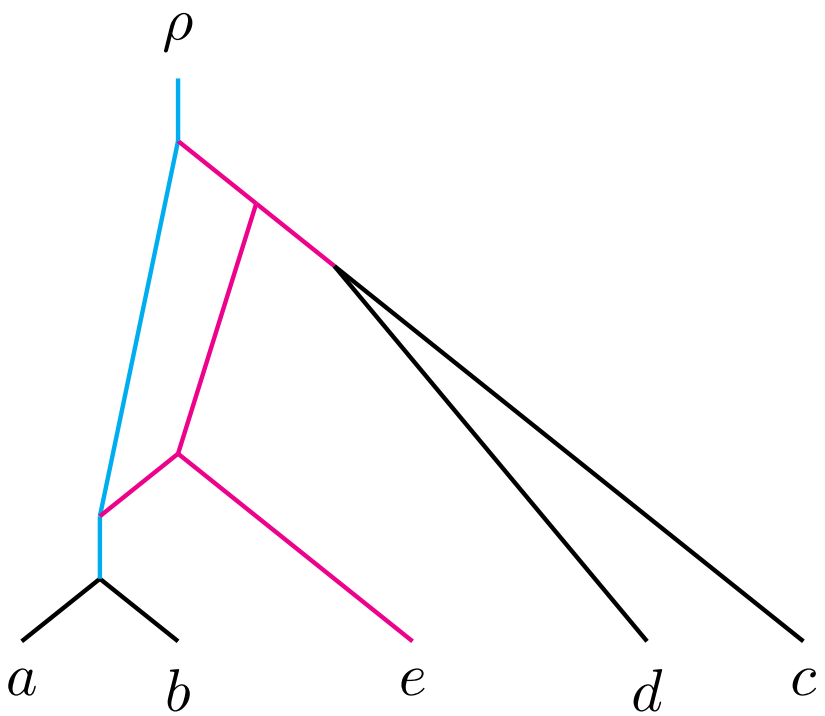


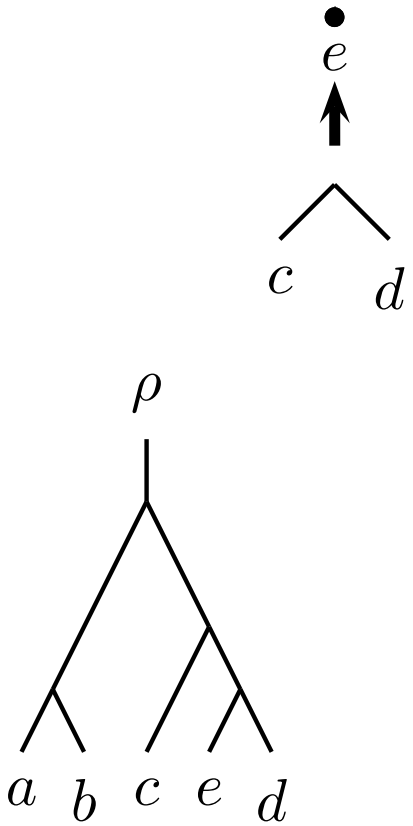
Great, so what's the problem?



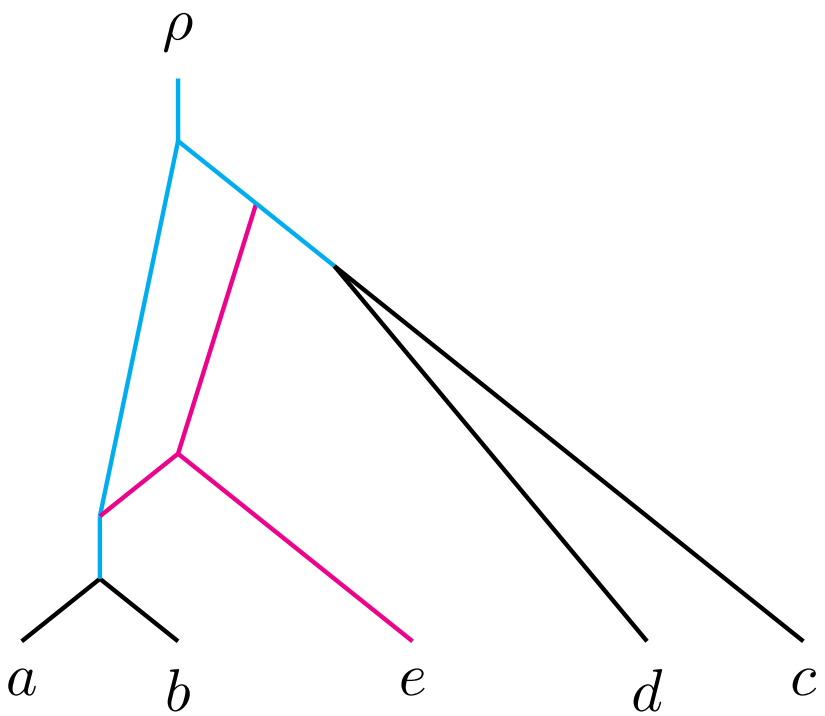


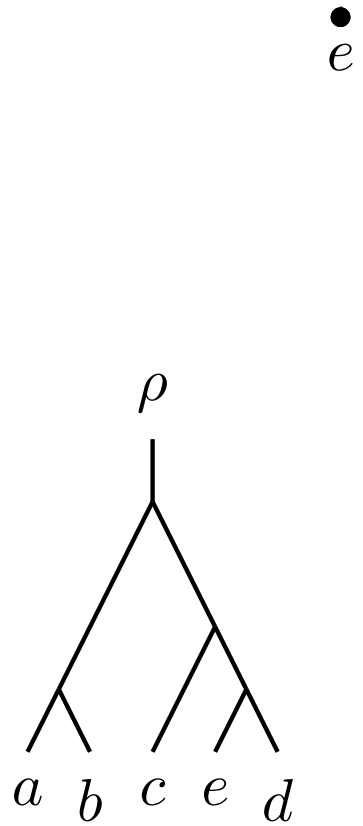
Great, so what's the problem?



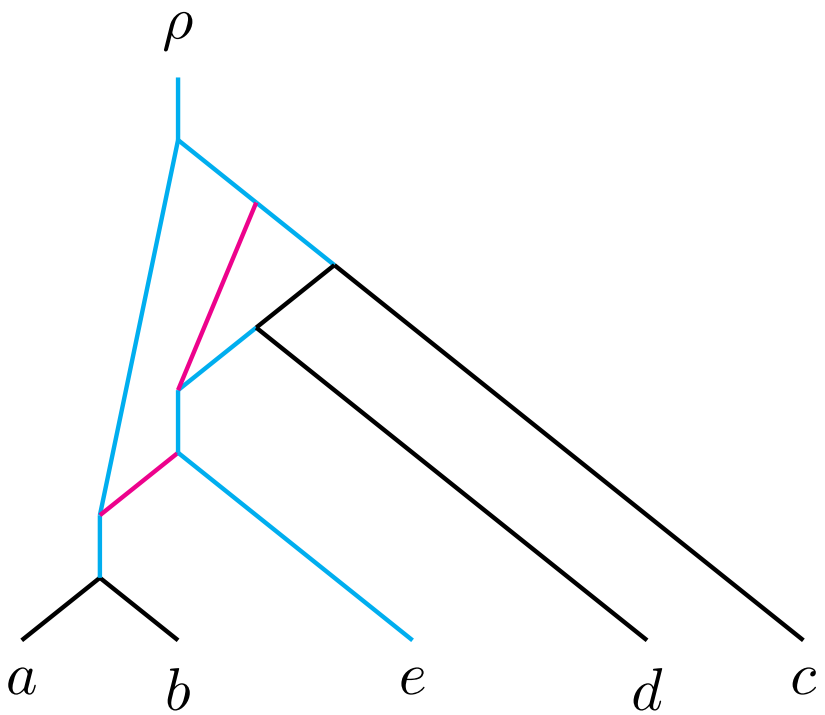


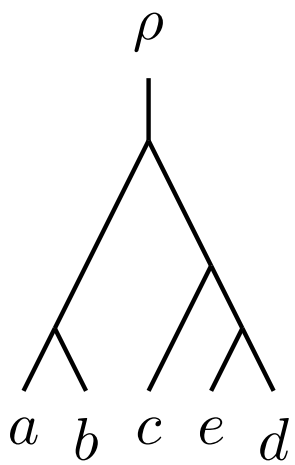
Great, so what's the problem?



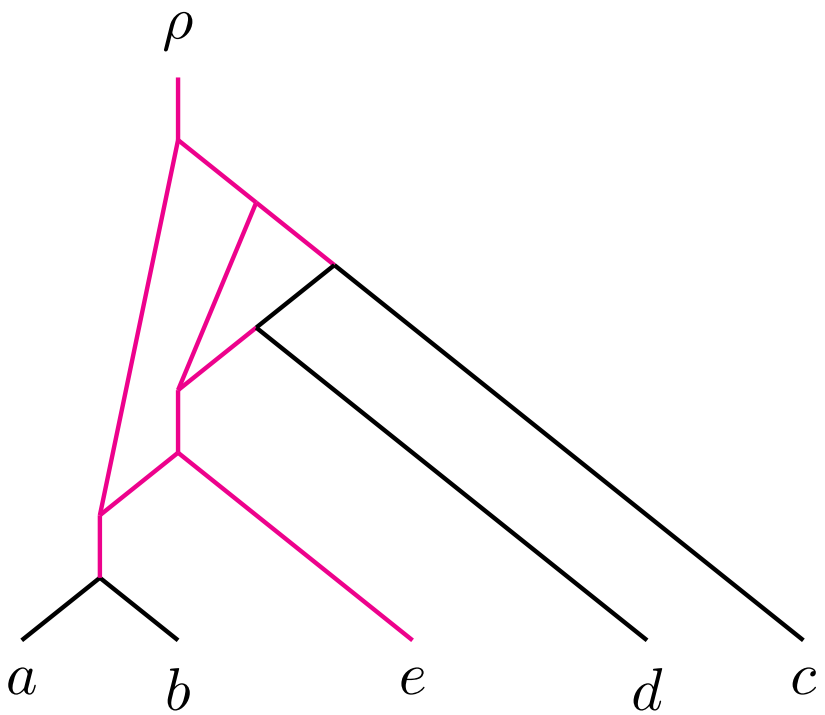


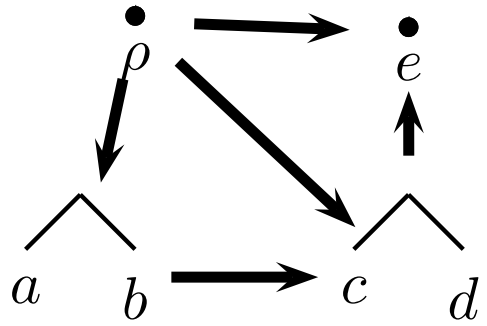
Great, so what's the problem?

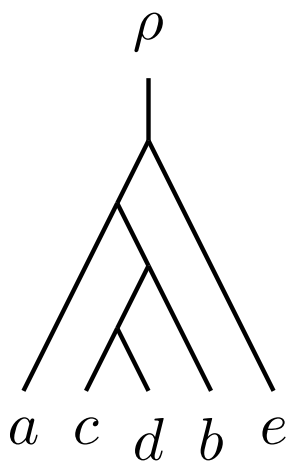




Great, so what's the problem?







Great, so what's the problem?

- Sometimes possible to add arcs in a way so that even if $\mathcal{T}_i \rightsquigarrow \mathcal{T}_j$ in $G_{\mathcal{F}}$ we have $\mathcal{T}_j \rightsquigarrow \mathcal{T}_i$ in $\mathcal{N}$.

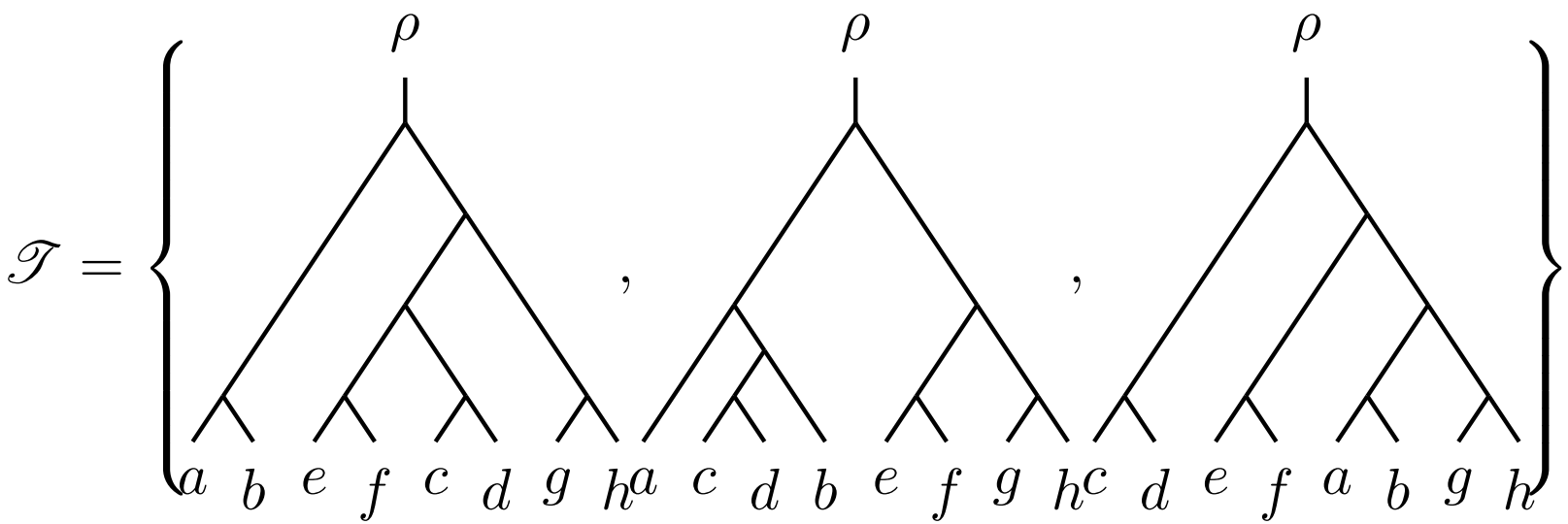
Great, so what's the problem?

- Sometimes possible to add arcs in a way so that even if $\mathcal{T}_i \rightsquigarrow \mathcal{T}_j$ in $G_{\mathcal{F}}$ we have $\mathcal{T}_j \rightsquigarrow \mathcal{T}_i$ in $\mathcal{N}$.
- Can be easily remedied by colouring any magenta arcs of paths terminating at the tree just added another colour, say fuchsia. Then we may add arcs that they can leave black, cyan, magenta or fuchsia coloured arcs but only enter magenta ones.

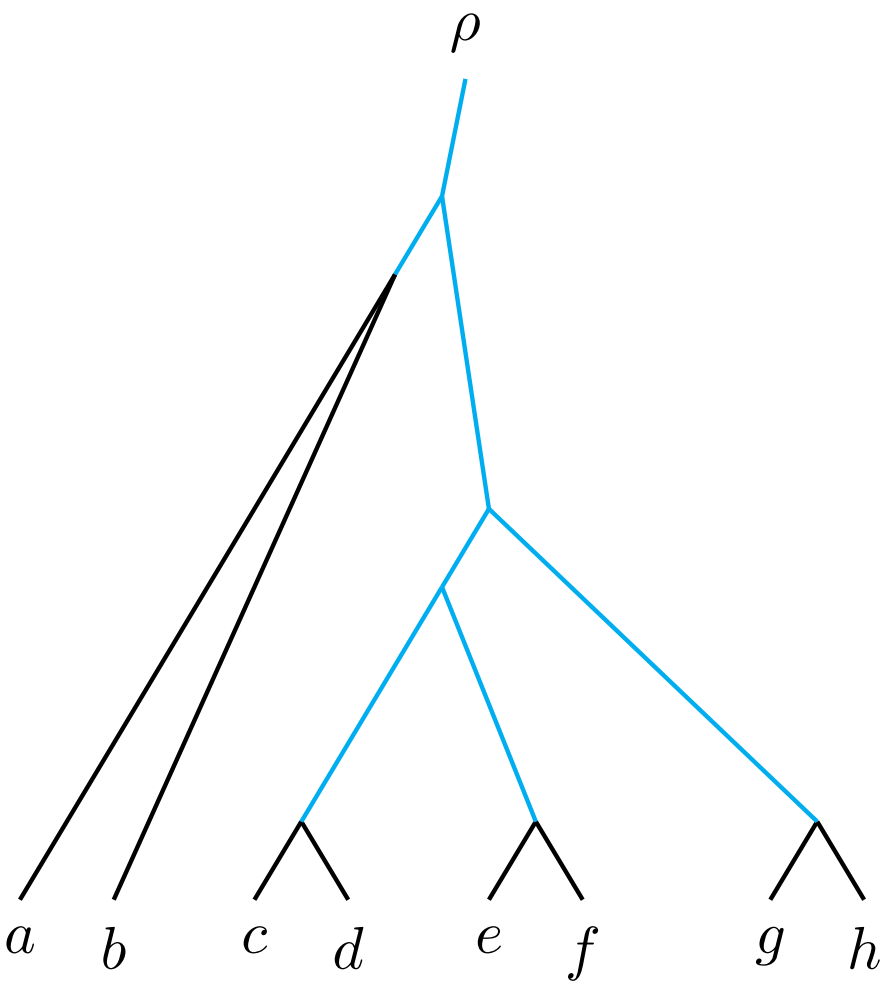
Great, so what's the problem?

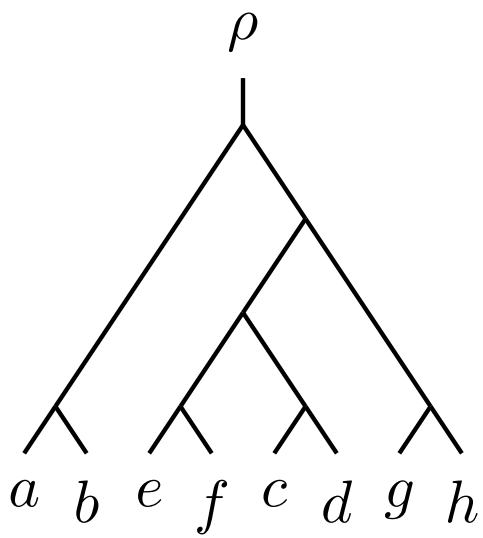
- Sometimes possible to add arcs in a way so that even if $\mathcal{T}_i \rightsquigarrow \mathcal{T}_j$ in $G_{\mathcal{F}}$ we have $\mathcal{T}_j \rightsquigarrow \mathcal{T}_i$ in $\mathcal{N}$.
- Can be easily remedied by colouring any magenta arcs of paths terminating at the tree just added another colour, say fuchsia. Then we may add arcs that they can leave black, cyan, magenta or fuchsia coloured arcs but only enter magenta ones.
- Can *still* get into positions where it's not possible to add a required arc without creating a cycle.

T_T

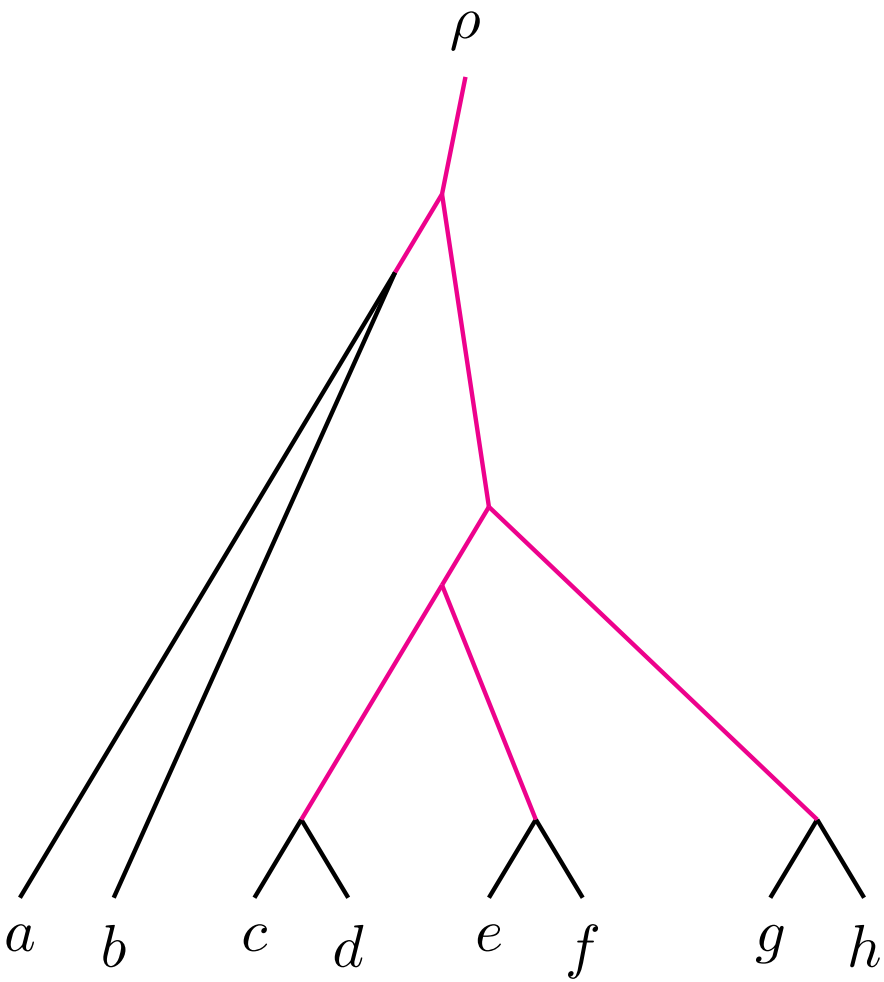


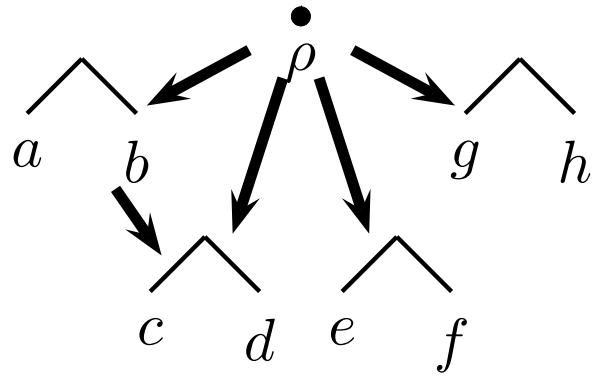
T_T

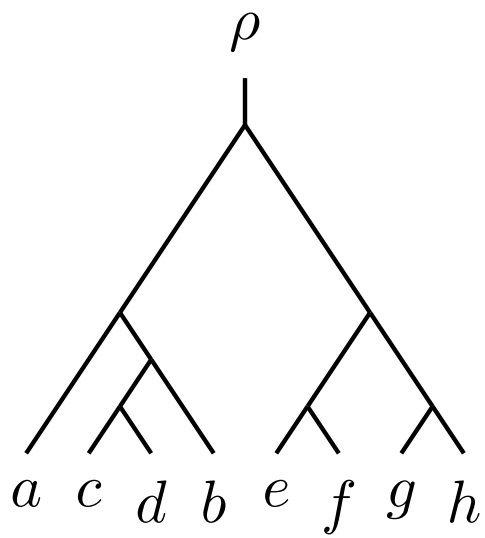




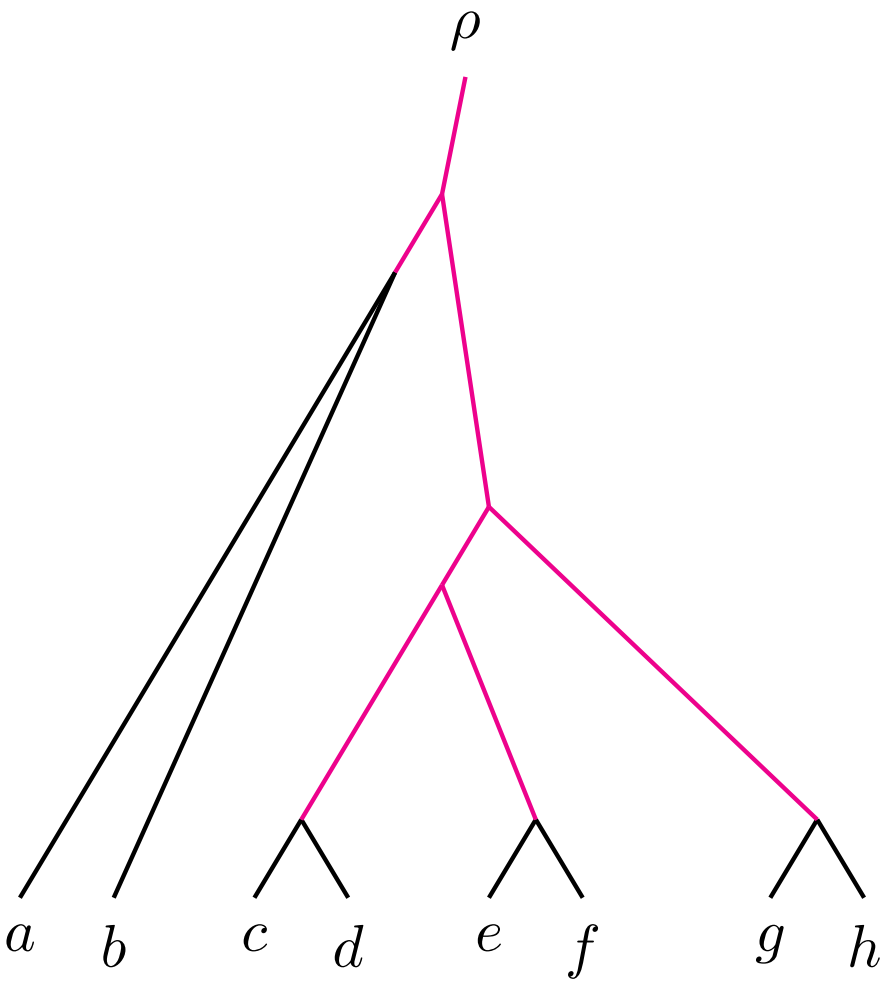
T_T

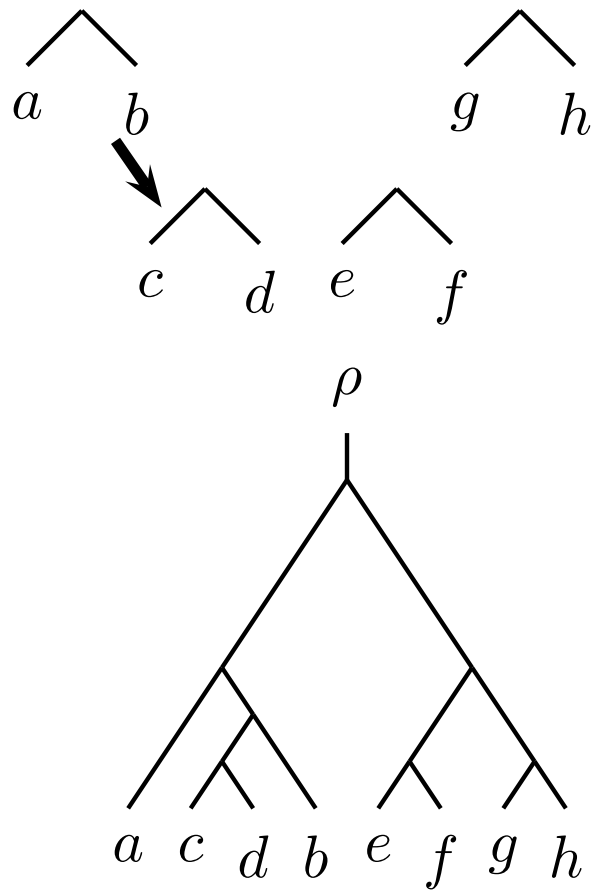




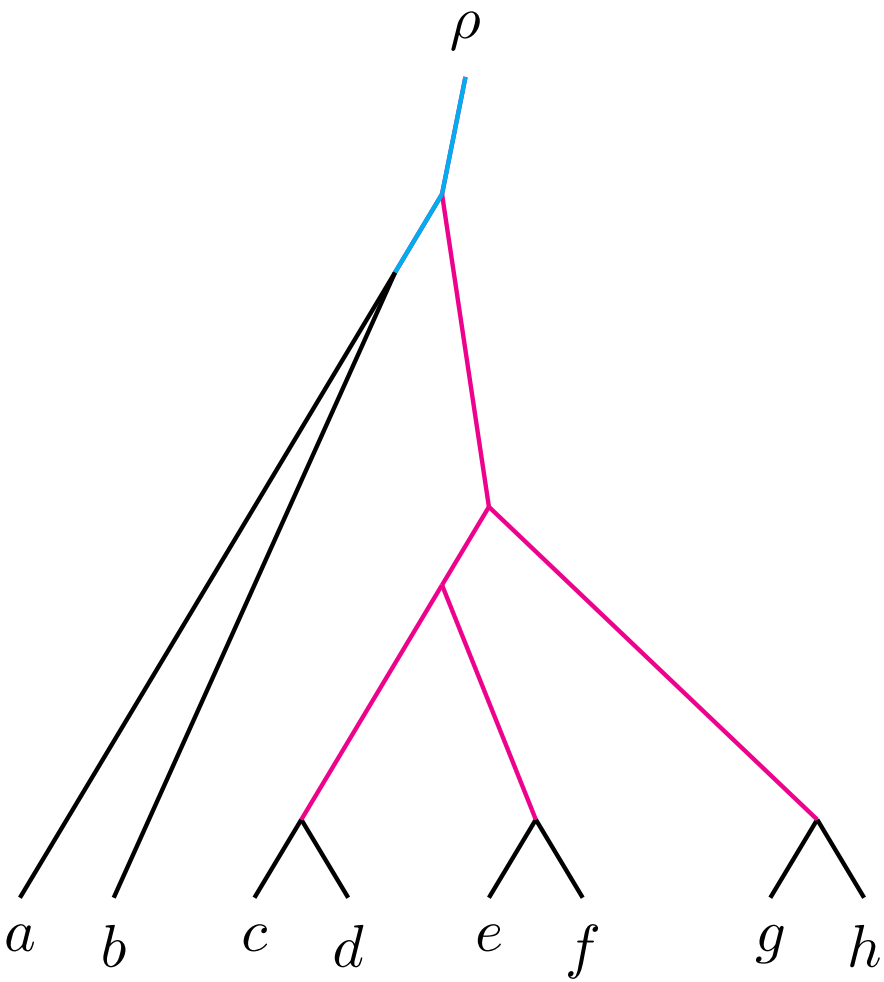


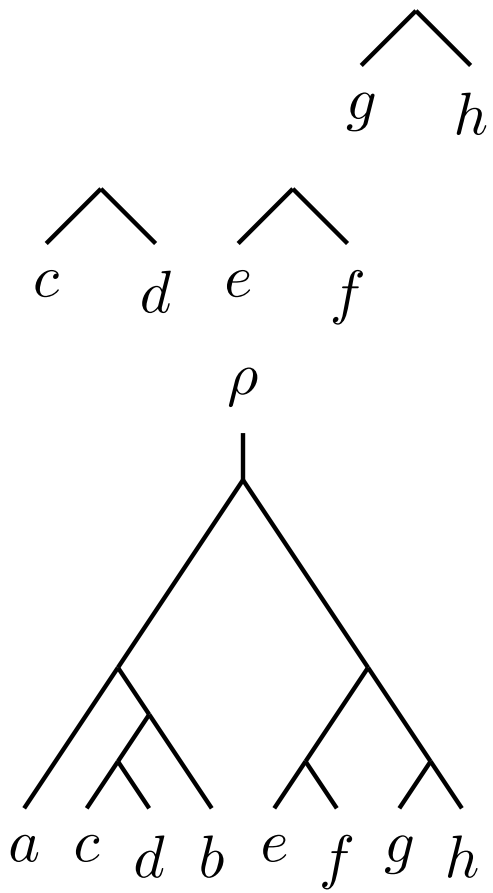
T_T



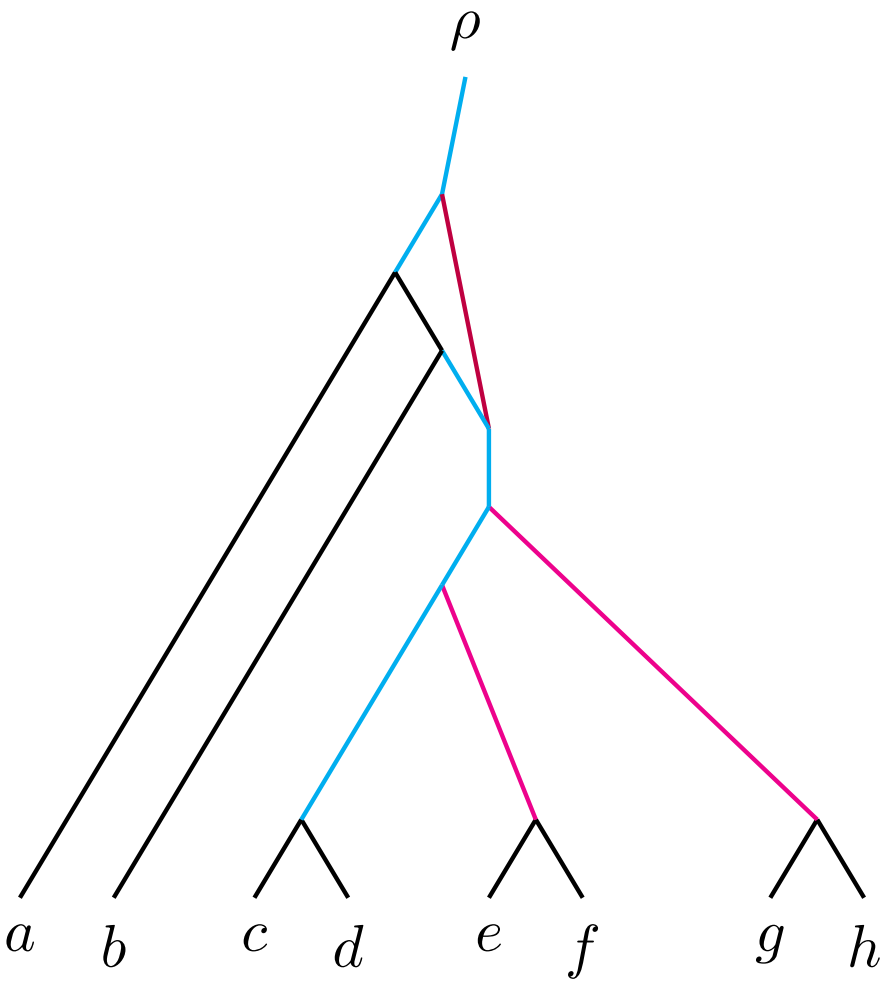


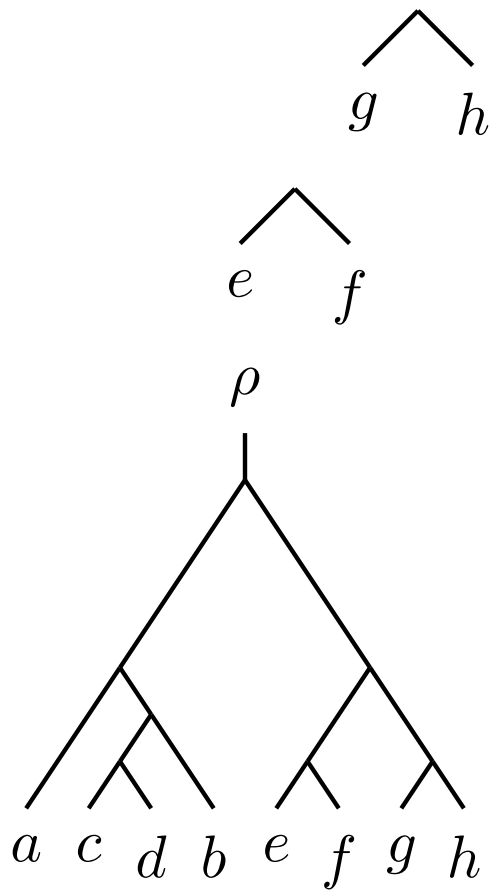
T_T



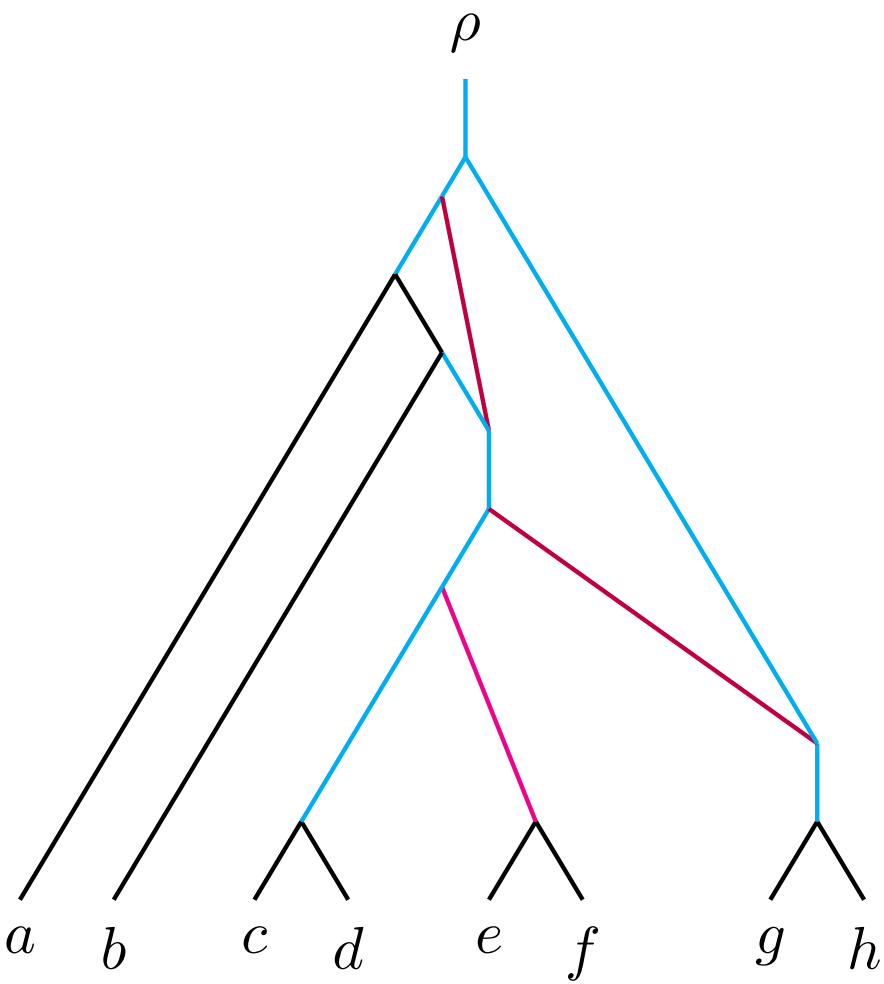


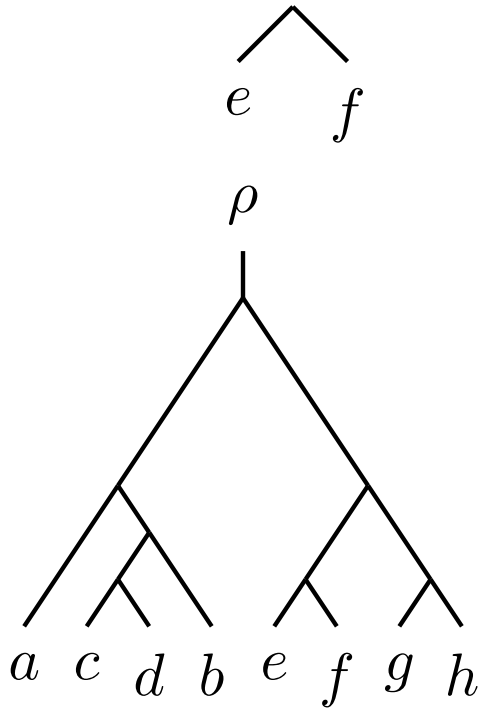
T_T



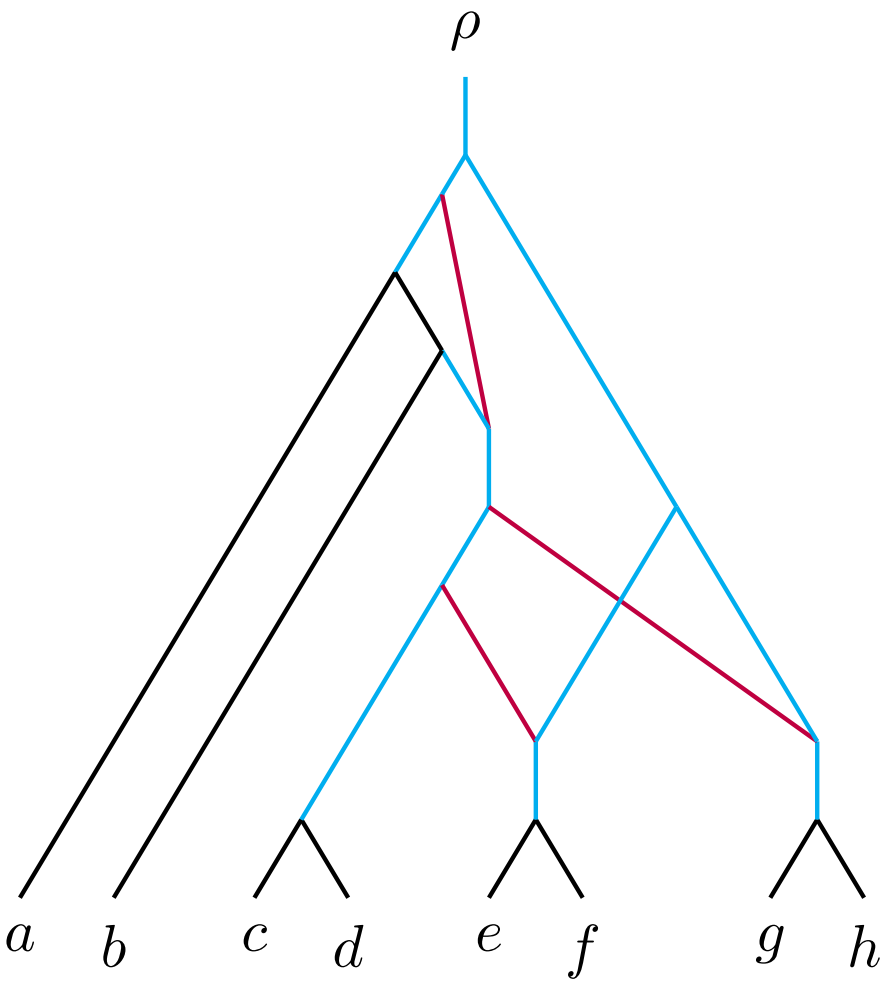


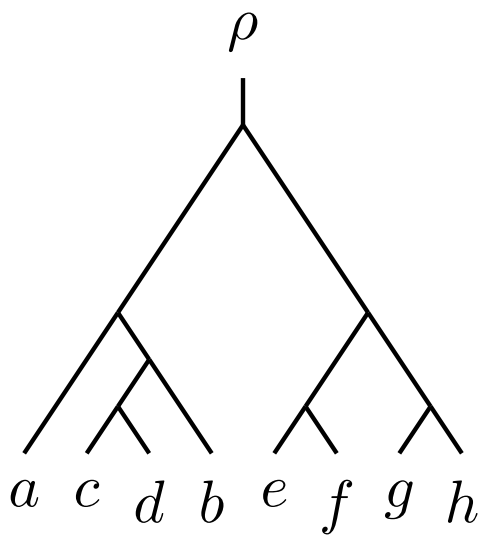
T_T



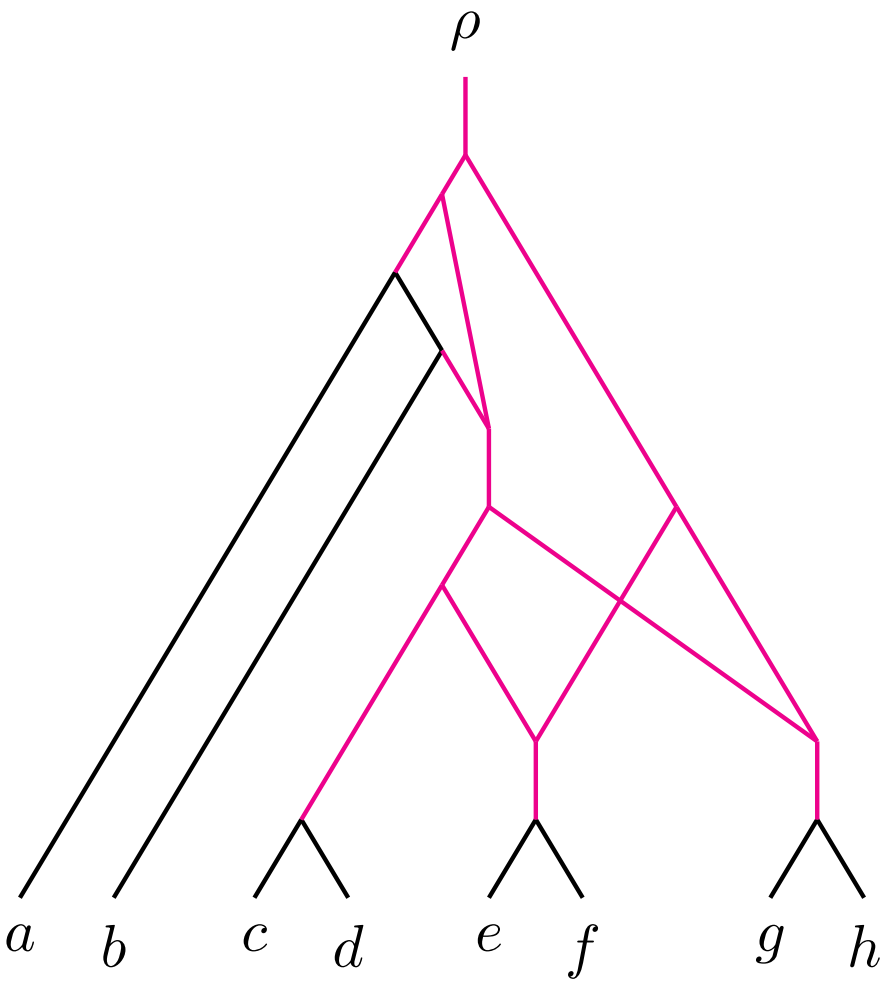


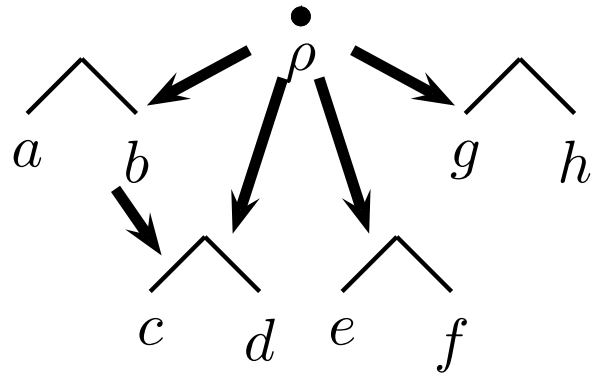
T_T

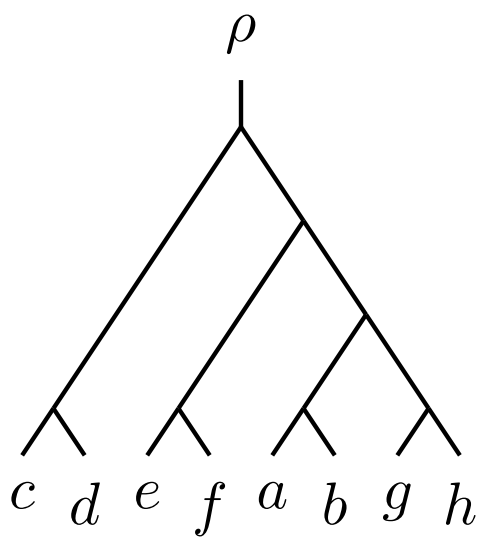




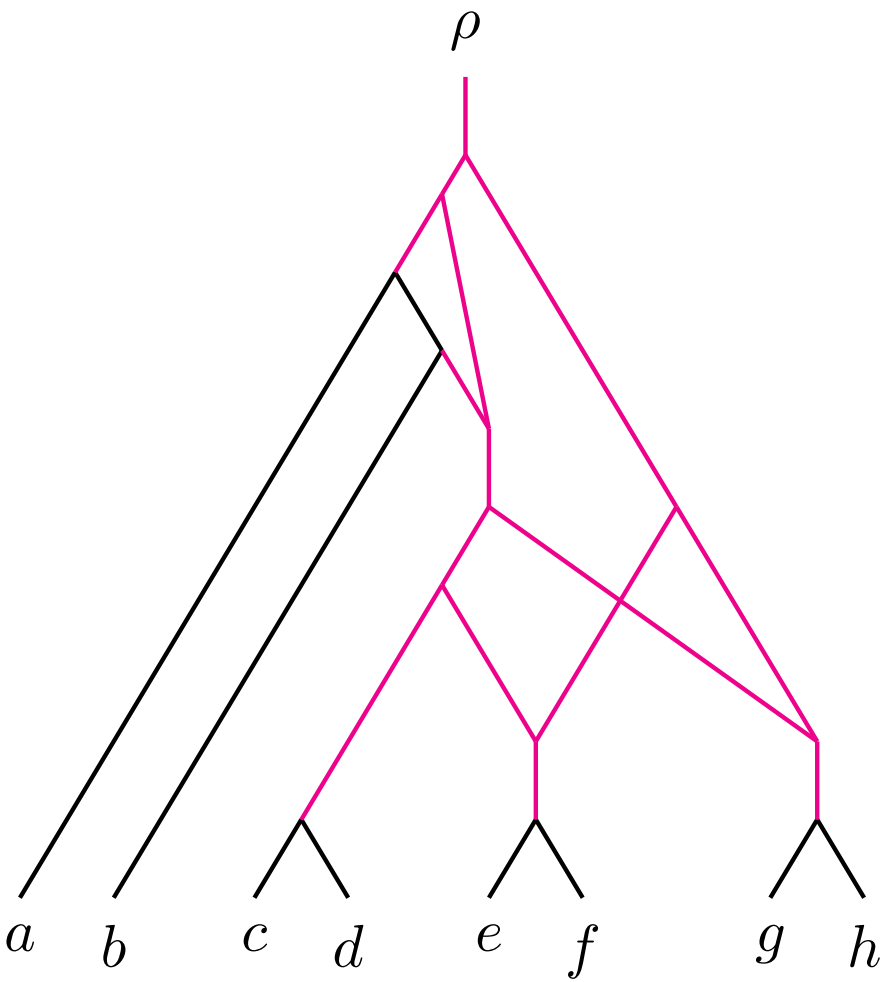
T_T

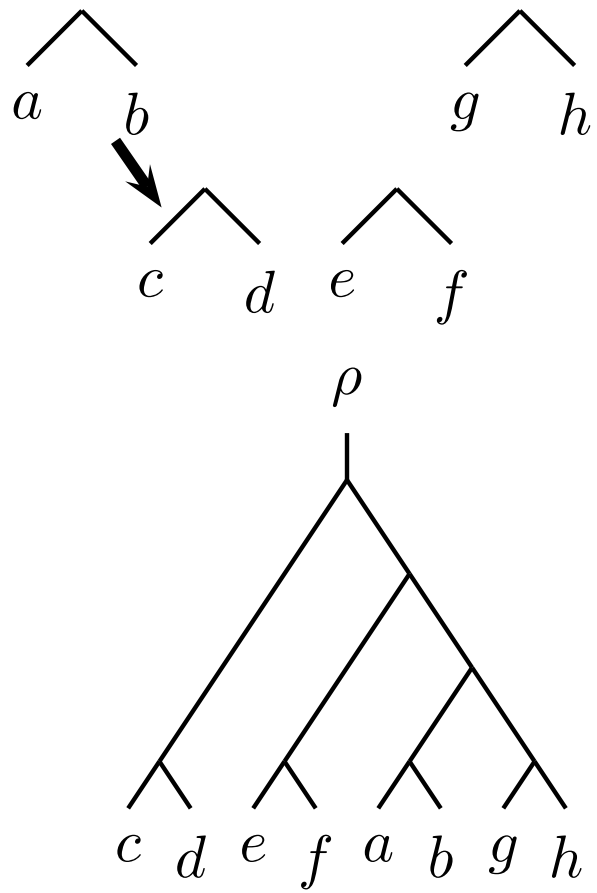




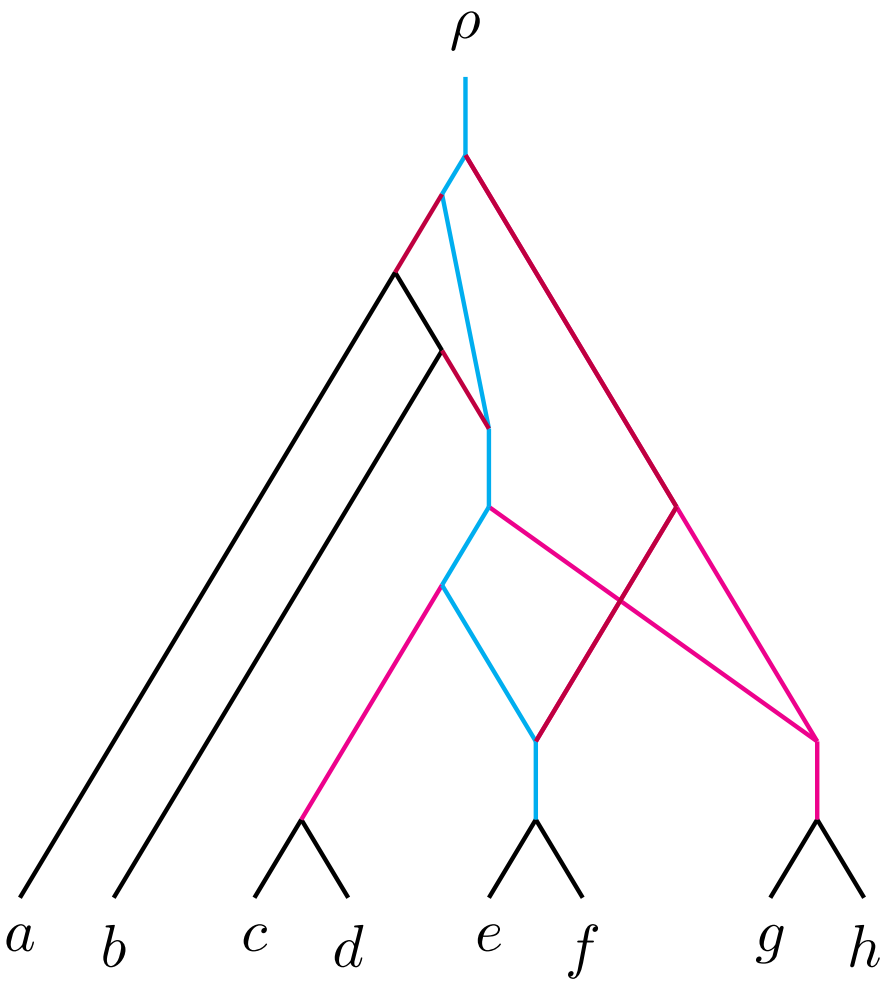


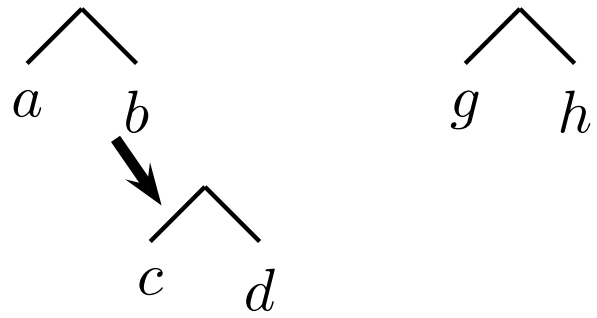
T_T

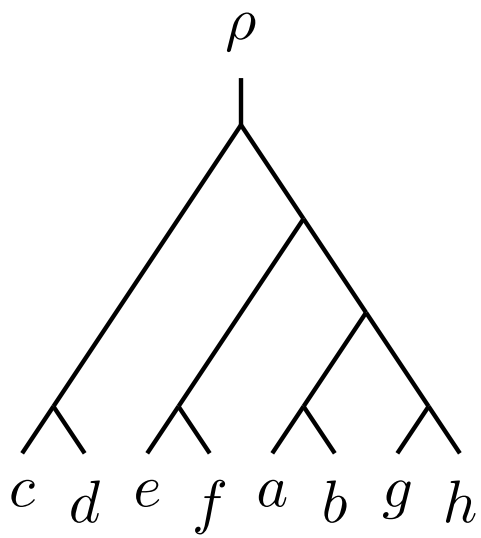




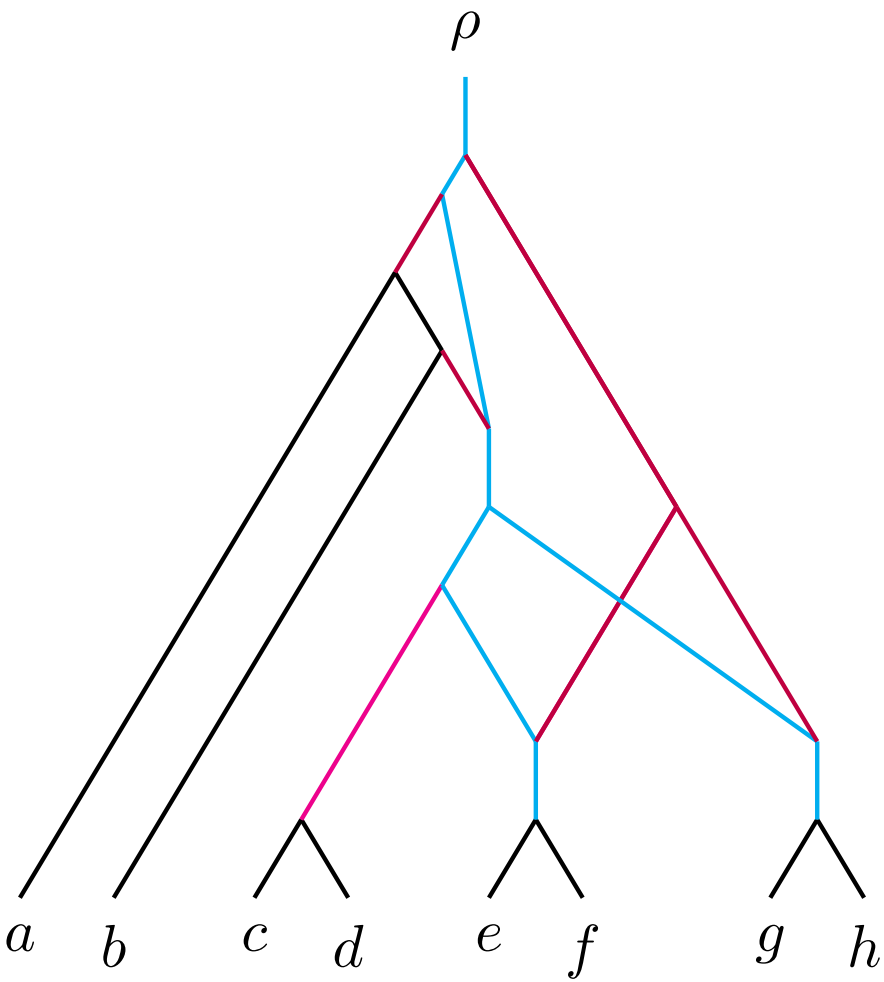
T_T

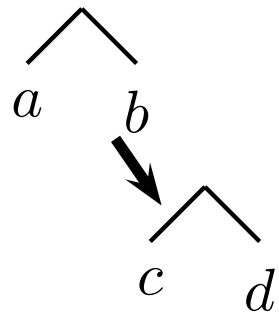


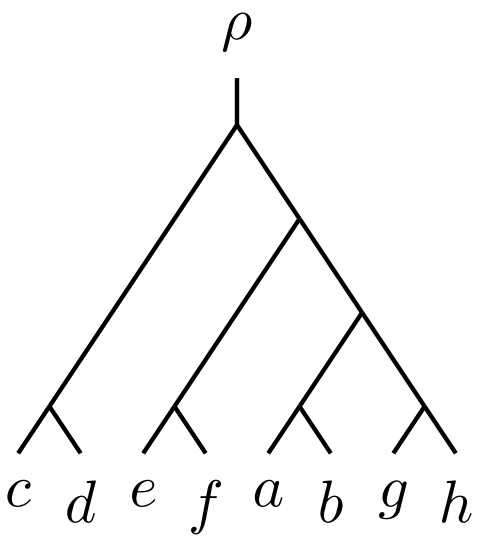




T_T





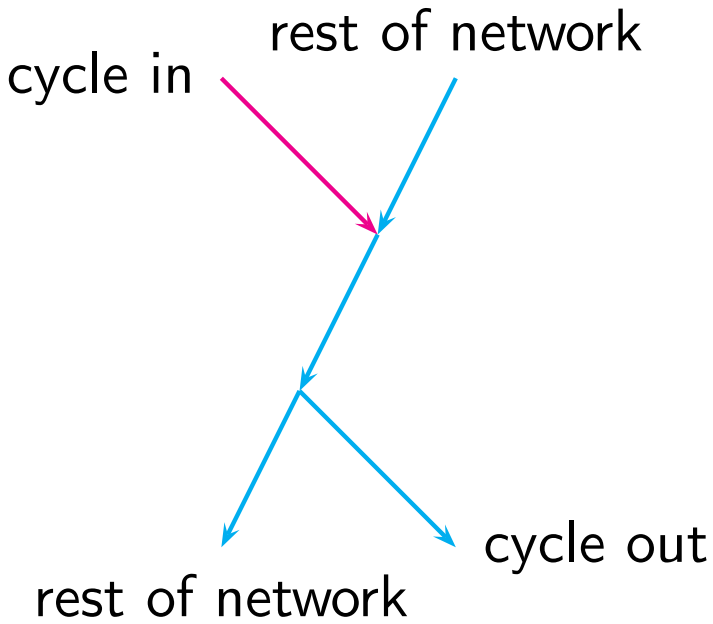


What do we want to do?

- Options are
  - avoid getting into situation
  - alter the phylogenetic network in some way that in polynomial time will alter it to another possible outcome.

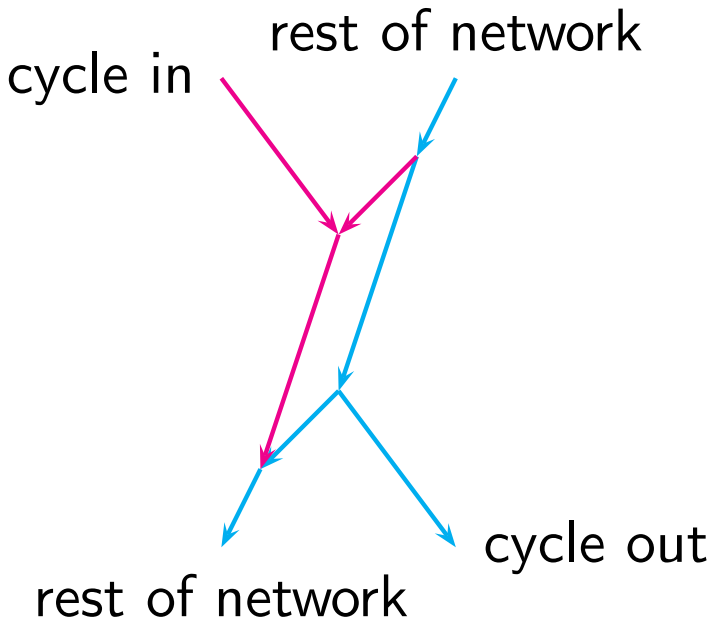
What do we want to do?

- Options are
  - avoid getting into situation
  - alter the phylogenetic network in some way that in polynomial time will alter it to another possible outcome.



What do we want to do?

- Options are
  - avoid getting into situation
  - alter the phylogenetic network in some way that in polynomial time will alter it to another possible outcome.



Under what conditions can we not use this 'pull-apart' operator?

- As per our aim of the algorithm that the forest induced by $\mathcal{N}$ is a superforest of the provided forest $\mathcal{F}$ we are not permitted to do this operator to black arcs.

Under what conditions can we not use this 'pull-apart' operator?

- As per our aim of the algorithm that the forest induced by $\mathcal{N}$ is a superforest of the provided forest $\mathcal{F}$ we are not permitted to do this operator to black arcs.
- This means that there may be no path that creates such a cycle that passes through black arcs after the most recent hybridisation event.

Under what conditions can we not use this 'pull-apart' operator?

- As per our aim of the algorithm that the forest induced by $\mathcal{N}$ is a superforest of the provided forest $\mathcal{F}$ we are not permitted to do this operator to black arcs.
- This means that there may be no path that creates such a cycle that passes through black arcs after the most recent hybridisation event.
- Can show contradiction arises with earlier addition of fuchsia arcs if we assume we have black arcs in irritating places.

Under what conditions can we not use this 'pull-apart' operator?

- As per our aim of the algorithm that the forest induced by $\mathcal{N}$ is a superforest of the provided forest $\mathcal{F}$ we are not permitted to do this operator to black arcs.
- This means that there may be no path that creates such a cycle that passes through black arcs after the most recent hybridisation event.
- Can show contradiction arises with earlier addition of fuchsia arcs if we assume we have black arcs in irritating places.
- Thus, never.

What next?

- Need to define explicit algorithm for identifying the sets between which arcs are added.

What next?

- Need to define explicit algorithm for identifying the sets between which arcs are added.
- Extend to weighted trees, thus requiring some notion of a weighted forest.

What next?

- Need to define explicit algorithm for identifying the sets between which arcs are added.
- Extend to weighted trees, thus requiring some notion of a weighted forest.
- Extend to weighted networks.

What next?

- Need to define explicit algorithm for identifying the sets between which arcs are added.
- Extend to weighted trees, thus requiring some notion of a weighted forest.
- Extend to weighted networks.

Other possible uses:

What next?

- Need to define explicit algorithm for identifying the sets between which arcs are added.
- Extend to weighted trees, thus requiring some notion of a weighted forest.
- Extend to weighted networks.

Other possible uses:

- Enumerate through all networks that display $\mathscr{T}$ and induce a forest that is a superforest of a provided $\mathcal{F}$

What next?

- Need to define explicit algorithm for identifying the sets between which arcs are added.
- Extend to weighted trees, thus requiring some notion of a weighted forest.
- Extend to weighted networks.

Other possible uses:

- Enumerate through all networks that display $\mathscr{T}$ and induce a forest that is a superforest of a provided $\mathcal{F}$
- Thus give a fairly brute force algorithm to calculate the hybridisation number of sets of trees of arbitrary size.

Open Questions

Question

In the two tree case the size of the maximum acyclic agreement forest is one greater than the hybridisation number of the set of trees. If given more than two trees is there any similar relation between the networks of smallest hybridisation number that displays the trees and maximum acyclic agreement forests? If not is there a relation between it and the maximal acyclic agreement forests?

Acknowledgements

- Mike Hendy, University of Otago
- Barbara Holland, University of Tasmania
- David Bryant, University of Otago
- Katharina Huber, University of East Anglia
- Vincent Moulton, University of East Anglia

- Otago University
- Massey University

- Marsden Grant